In compliance with the
Canadian Privacy Legislation
some supporting forms
may have been removed from
this dissertation.

While these forms may be included
in the document page count,
their removal does not represent
any loss of content from the dissertation.

University of Alberta

Working Memory and Syllogistic Reasoning in Computer-Based Testing

by

Matthew M. Gushta

A thesis submitted to the Faculty of Graduate Studies and Research in partial fulfillment of the requirements for the degree of Master of Education

Department of Educational Psychology

Edmonton, Alberta

Fall, 2003

National Library
of Canada

Acquisitions and
Bibliographic Services

395 Wellington Street
Ottawa ON K1A 0N4
Canada

Bibliothèque nationale
du Canada

Acquisisitons et
services bibliographiques

395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file Votre référence
ISBN: 0-612-87606-3
Our file Notre référence
ISBN: 0-612-87606-3

The author has granted a non-exclusive licence allowing the National Library of Canada to reproduce, loan, distribute or sell copies of this thesis in microform, paper or electronic formats.

The author retains ownership of the copyright in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.

L'auteur a accordé une licence non exclusive permettant à la Bibliothèque nationale du Canada de reproduire, prêter, distribuer ou vendre des copies de cette thèse sous la forme de microfiche/film, de reproduction sur papier ou sur format électronique.

L'auteur conserve la propriété du droit d'auteur qui protège cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou aturement reproduits sans son autorisation.

Canadä

University of Alberta

Library Release Form

Name of Author: Matthew M. Gushta

Title of Thesis: Working Memory and Syllogistic Reasoning in Computer-Based Testing

Degree: Master of Education

Year this Degree Granted: 2003

Permission is hereby granted to the University of Alberta Library to reproduce single copies of this thesis and to lend or sell such copies for private, scholarly, or scientific research purposes only.

The author reserves all other publication and other rights in association with the copyright in the thesis, and except as herein before provided, neither the thesis nor any substantial portion thereof may be printed or otherwise reproduced in any material form whatever without the author's prior written permission.

August 27, 2003

University of Alberta

Faculty of Graduate Studies and Research

The undersigned certify that they have read, and recommend to the Faculty of Graduate Studies and Research for acceptance, a thesis entitled **Working Memory and Syllogistic Reasoning in Computer-Based Testing** submitted by **Matthew M. Gushta** in partial fulfillment of the requirements for the degree of **Master of Education**.

Dr. Jacqueline Leighton (Supervisor)

Dr. Mark Gierl

Dr. Connie Varnhagen

August 14, 2003

ABSTRACT

Equivalency of paper-based tests (P&Ps) and computer-based tests (CBTs) has been examined for a number of years (i.e., Mead & Drasgow, 1993; Epstein & Klinkenberg, 2001). In an attempt to illuminate any differences between P&P and CBT performance this study analyzed 209 examinees' responses to twenty-four categorical syllogisms at both the test and item level, incorporating working memory theory in accounting for any differential performance. While cross-modal correlations were high ($r = 0.97$, $p<.05$), significant differences were found between P&P and CBT performance, as P&P administration resulted in increased performance ($F_{1,205} = 4.58$, $p<.05$). This difference is further explained given examinees' visuo-spatial ability and access to an external memory aid. Performance on difficult items or by low ability examinees was facilitated by the P&P and aid-provided conditions; difficult items and high ability examinees performed better within CBT and aid-prohibited conditions – evidence of the *Expertise Reversal Effect*.

DEDICATION

This thesis is dedicated to my father, Maurice Gushta. Through his words I have been encouraged to achieve all that I desire; by his actions I know what it is to persevere.

ACKNOWLEDGEMENTS

I would like to recognize the immeasurable support and guidance provided by the faculty members in the Centre for Research in Applied Measurement and Evaluation: my supervisor, Dr. Jacqueline Leighton, and Drs. Mark Gierl and Todd Rogers; the freedom offered by Dr. Connie Varnhagen; and the inspiration fostered by Dr. Stanley Varnhagen.

TABLE OF CONTENTS

CHAPTER I - BACKGROUND 1

OBJECTIVE OF THE STUDY 1

INTRODUCTION 2

COGNITION IN TESTING 4

STATISTICAL EQUIVALENCE FOR TEST COMPARABILITY 5

COGNITIVE COMPARABILITY 11

SYLLOGISTIC REASONING 16

PAPER FOLDING TEST 20

CURRENT STUDY 21

RESEARCH QUESTIONS 22

CHAPTER II - METHODS 23

PARTICIPANTS 23

DESIGN 23

PROCEDURE AND MATERIALS 23

HYPOTHESES 26

Hypothesis 1: Mode of Administration 26

Hypothesis 2: External Memory Aid 26

Hypothesis 3: Interaction 27

Hypothesis 4: VSSP 27

CHAPTER III - RESULTS **28**

Demographics 28

Item Correlation 29

Test-Level Analyses 29

Item-Level Analyses 31

CHAPTER IV – CONCLUSION **41**

Summary 41

Discussion 45

FOOTNOTES **46**

REFERENCES **47**

TABLES **57**

FIGURES **79**

Figure Captions 79

APPENDICES **103**

Appendix A: Syllogisms 104

Appendix B: Paper Folding Test Example 106

Appendix C: Sample Instrument 107

Appendix D: Item Characteristic Curves 112

LIST OF TABLES

Table 01 Syllogistic Categories 57

Table 02 Demographic Proportions of Experimental Sample 58

Table 03 Five-Point Computer Self-Assessment 59-60

Table 04 Item p-Values Across Experimental Conditions 61

Table 05 Intercorrelations Between Item p-Values Across Conditions 62

Table 06 Syllogistic Performance 63

Table 07 ANOVA for Total Syllogistic Performance 64

Table 08 Syllogistic Performance Across Modes and Memory Aid Use 65

Table 09 ANOVA for Syllogistic Performance Across Modes and Memory Aid Use 66

Table 10 Paper Folding Test Performance Across Modes of Administration 67

Table 11 Syllogistic Performance Across Modes and Memory Aid Given Visuo-Spatial Ability 68

Table 12 ANOVA for Syllogistic Performance Across Modes and Memory Aid Given Visuo-Spatial Ability 69

Table 13 Syllogistic Performance Across Modes and Memory Aid Use Given Visuo-Spatial Ability 70

Table 14 ANOVA for Syllogistic Performance Across Modes and Memory Aid Use Given Visuo-Spatial Ability 71

Table 15 Beta Values Across Experimental Conditions 72

Table 16 Beta Values Across Mode of Administration and Use of Memory Aid 73

Table 17 Beta Values Across Experimental Conditions According to Low Visuo-Spatial Ability 74

Table 18 Beta Values Across Mode of Administration and Use of Memory Aid According to Low Visuo-Spatial Ability 75

Table 19 Beta Values Across Experimental Conditions According to High Visuo-Spatial Ability 76

Table 20 Beta Values Across Mode of Administration and Use of Memory Aid According to High Visuo-Spatial Ability 77

Table 21 Direction of DIF Evidenced Across Experimental Conditions, Ability, and Syllogism Type 78

LIST OF FIGURES

Figure 01 Syllogistic Performance According to Ability and Memory Aid Usage..... 82

Figure 02 Item Characteristic Curve (ICC) generated by Testgraf..... 83

Figure 03 Sample DIF Graph for Overall Ability and Multiple-Model Syllogisms..... 84

Figure 04 Overall Ability and One-Model Syllogisms..... 85

Figure 05 Overall Ability and Multiple-Model Syllogisms..... 86

Figure 06 Overall Ability and No Valid Conclusion Syllogisms 87

Figure 07 Low Ability and One-Model Syllogisms 88

Figure 08 Low Ability and Multiple-Model Syllogisms 89

Figure 09 Low Ability and No Valid Conclusion Syllogisms..... 90

Figure 10 High Ability and One-Model Syllogisms..... 91

Figure 11 High Ability and Multiple-Model Syllogisms..... 92

Figure 12 High Ability and No Valid Conclusion Syllogisms 93

Figure 13 Memory Aid Use by Overall Ability and One-Model Syllogisms..... 94

Figure 14 Memory Aid Use by Overall Ability and Multiple-Model Syllogisms..... 95

Figure 15 Memory Aid Use by Overall Ability and No Valid Conclusion Syllogisms 96

Figure 16 Memory Aid Use by Low Ability and One-Model Syllogisms 97

Figure 17 Memory Aid Use by Low Ability and Multiple-Model Syllogisms 98

Figure 18 Memory Aid Use by Low Ability and No Valid Conclusion Syllogisms..... 99

Figure 19 Memory Aid Use by High Ability and One-Model Syllogisms..... 100

Figure 20 Memory Aid Use by High Ability and Multiple-Model Syllogisms..... 101

Figure 21 Memory Aid Use by High Ability and No Valid Conclusion Syllogisms 102

Figure B1 Paper Folding Test example 106

Figure C1 Sample instrument page one 107

Figure C2 Sample instrument page two 108

Figure C3 Sample instrument page three 109

Figure C4 Sample instrument page four 110

Figure C5 Sample instrument page eight 111

Figure D1 Item Characteristic Curves (ICCs) for the twenty-four syllogisms 112

CHAPTER I - BACKGROUND

Objective of the Study

The purpose of the present study was to investigate the *equivalency* or *comparability* between distinct modes of testing: computer-based tests (CBTs) and paper-and-pencil based tests (P&Ps), with attention paid to the influence of specific cognitive factors. The rationale for the study has come from a review of the literature, which suggests that the equivalency between computer- and paper-and-pencil based tests is still questioned in the psychometric community. In the hope of gaining some understanding of the equivalency between these two distinct modes of testing, the present study manipulates mode of testing within a controlled cognitive domain. One cognitive factor, working memory and specifically the visuo-spatial sketchpad (VSSP) function, was examined for its mediating role in the equivalency of CBTs and P&Ps. Previous research has shown that the provision of such aids as note pads, serving to reduce the load on working memory, can help to improve performance on cognitive tasks. The domain of syllogistic reasoning was chosen to investigate the equivalency of CBTs and P&Ps because the task factors that influence syllogistic reasoning performance are well known in the cognitive psychological literature.

Introduction

Technological capability is ever increasing, affording educators, administrators, and test developers, among many others, ever-expanding potential with respect to the manner of test delivery and type of content available for students and examinees. Specifically, these advancements have directly translated into the development of computer-based testing applications for both classroom and large-scale testing programs (Bennett, 1998; Epstein & Klinkenberg, 2001). The simplest method for computerization of assessment instruments is the *linear* CBT. Similar to their P&P counterparts in most respects other than medium of administration, linear CBTs present items according to a predetermined order. This design stands in contrast to the advent of computerized-adaptive tests (CATs), the form most often studied in CBT research. As a result of technological and psychometric advances, CATs are able to utilize modern *item response theory* (see Hambleton, Swaminathan, & Rogers, 1991, for a review of IRT). In CAT, each new item is presented based on an examinee's estimated ability given his or her performance on previously presented items (for a review, see Wainer, 2000). The term *adaptive* is derived from the fact that these tests are tailored to each examinee's ability yielding uniquely constructed instruments for each administration.

Even with such technical advances, the shift to computer-based testing (or adaptive testing, which introduces other, more complex issues) is being made gradually, requiring that *equivalence* across modes of administration (i.e., CBT and P&P) be shown. Equivalency of CBT and P&P administration needs to be determined given that exams are being offered in both forms concurrently, with a number of exams moving to entirely

computerized systems (e.g., Graduate Records Exam [GRE]; Test of English as a Foreign Language [TOEFL]; Armed Services Vocational Aptitude Battery [ASVAB]).

Equivalence is the term used in the literature to reflect the notion of interchangability or, more appropriately, parallelism across modes of administration (Lord & Novick, 1968). Thus far, equivalence has been accepted (or contested) on the grounds of statistical evidence, defined as the attainment of strong correlations and similar factor structures across modes of administration (Mead & Drasgow, 1993). However, the notion of "equivalence" is somewhat misleading given that the assertion of parallel cognitive experiences cannot be made even if statistical equivalence is obtained across modes of administration. Hence, although the term "equivalence" is used in the literature to refer to parallelism across modes, it is more accurate to talk about "comparability" of tests across modes of administration.

Aside from statistical indices, assessing the comparability of tests across modes of administration requires that attention also be paid to issues of human-computer interface (Booth, 1991, 1998a, 1998b; Honaker, 1988). Complementary to psychometric and technological innovations in assessment, theoretical advances from the cognitive psychological domain are beginning to see application through the dissemination and slow implementation of *cognitively principled design* (Bennett, 1999). Cognitively principled design refers to the application of cognitive science research to the development and interpretation of performance assessment instruments that have historically been based on behavioural principles. By Bennett's (1998) definition, it is expected that the end of the era of "first-generation" CBTs (automation of assessments without reconceptualization or redesign) is fast approaching. Moving next into the

development of "second-generation" CBTs, incorporating varieties of item and response formats developed with cognitive principles in mind (Bennett, 1998, p.5). Ultimately, generation "R" assessments (Bennett, 1998, p. 11) will be the result of the *reinvention* of computer-based testing – seamlessly embedding assessment within instruction, serving to simultaneously assess and provide feedback as appropriate for the task, informed by models and theories of subject-matter expertise.

Through the application of cognitive psychology to measurement and assessment issues, psychometricians, test developers, and instructors may be afforded a clearer picture of computer-based performance. Although cognitive theory is still not completely represented in assessment (Booth, 1998), gradually, knowledge about cognitive processing and performance on cognitive tasks will be incorporated (Bennett, 1998, 1999). Application of cognitive psychology to measurement and assessment issues offers the potential to further explain and inform the development of CBTs, and clearly illustrate issues regarding the equivalence or comparability between CBTs and P&Ps.

Cognition in Testing

Sternberg (1984) states, "cognitive psychology has provided a valuable complementary way of investigating pretty much the same constructs psychometricians have been studying all along" (p. 49). Differences between cognitive and psychometric score interpretations, however, lie in the fact that the psychometric approach provides mere artefacts of performance in the form of a test score. A cognitive approach allows for examination of actual processes underlying test performance, which are important to examine if testing is to inform instruction and support valid inferences (Bennett, 1999; Huff & Sireci, 2001; Snow & Lohman, 1989; Sternberg, 1984). Shifts in the purpose of

assessment, toward informing instructional practice, and technological and theoretical advancements in computing and measurement and cognitive domains, have allowed and facilitated the merger of the two disciplines (Pellegrino, et al., 1999; Sternberg, 1984). The confluence of educational measurement and cognitive science can be seen through a matrix of educational activities and basic cognitive principles; assessments are adequate only when they measure performance according to content as well as accounting for cognitive processing and ability (Elshout, 1985; Snow & Lohman, 1989).

Correlational approaches have been used extensively in both educational measurement and cognitive psychology (i.e., Mead & Drasgow, 1993; Sternberg, 1984). The basis of these methods, however, is simply the assessment of linearity and rank ordering across conditions. Well suited for classification purposes, the use of indices of correlation within cognitively informed test interpretation is inadequate as it only accounts for broad constructs (i.e., mathematical performance), typically represented by an indicator of overall performance. As the purpose of assessment shifts towards diagnosis and instructional remediation, more advanced methods must be employed if elementary cognitive factors (i.e., subtraction) are to be assessed and understood, not simply ranked.

Statistical Equivalence for Test Comparability

For the purpose of this study, I will define statistical equivalence as the provision of statistical evidence illustrating the absence of performance differences resulting from a comparison of computer- and paper-based test administration modes. It is important that such equivalence be shown because many exams are moving towards computerized administrations, co-existing with paper forms, or have been completely computerized

(e.g., GRE, TOEFL, ASVAB). In both instances, decisions being made based on results from CBTs must necessarily be compared to preceding or concurrent P&P results (McDonald, 2002; Wang & Kolen, 2001).

In the literature, the most common method for establishing statistical equivalence has been through correlational analyses. A meta-analysis conducted by Mead and Drasgow (1993) examined 159 cross-modal correlations reported in 28 studies comparing performance on the ASVAB, Differential Aptitude Tests (DAT), GRE, Multidimensional Aptitude Battery (MAB), Western Personnel Test (WPT), and others. Covering a wide range of content areas and research designs, including both within- and between-subjects designs, Mead and Drasgow (1993) reported an overall disattenuated correlation of .91 between computer- and paper-based exams.

While a number of studies have reported statistical equivalence of CBTs and P&Ps (Luecht, Hadadi, Swanson, & Case, 1998; Neuman & Baydoun, 1998; Segall & Moreno, 1999; Spray, Ackerman, Reckase, & Carlson, 1989), other studies comparing the effects of computerization on test performance have reported results that undermine the assumed equivalence of CBTs and P&Ps (Dimock & Cormier, 1991; Divgi & Stoloff, 1986; Fulcher, 1999; Greaud & Green 1986; Lee, Moreno, & Sympson, 1986; Noyes & Garland, 2003; Reinking & Schreiner, 1985; Van de Vijver & Harsveld, 1994), and others have reported mixed results (Mead & Drasgow, 1993; Russell, 1999; Russell & Haney, 1997; Steinberg, Thissen, & Wainer, 1990; Strong & Smith, 2002). Correlational studies include a wide range of values across modes, either approaching unity (e.g. Mead & Drasgow, 1993) or showing a weak relationship (e.g. Greaud, & Green, 1986). Furthermore, experimental analyses have shown evidence for differential performance at

the test or item level (Greaud, & Green, 1986) or have shown equivalence (Spray et al., 1989). Studies utilizing item response theory also provide contradictory results regarding cross-modal equivalence (Divgi & Stoloff, 1986; Luecht et al., 1998).

For example, Greaud and Green (1986) conducted a study examining cross-modal equivalence of computer- and paper-based forms of the Numerical Operations (NO) and Coding Speed (CS) subtests of the ASVAB using fifty college students. Each student completed two versions of each of the paper- and computer-based administrations for each subtest, for a total of eight tests per student. In the computer-based condition, students were given minimal editorial control – using pre-specified keys, students entered their answer to each question but could not skip, change, or later review items. Using the ratio of correct answers per minute, analyses were conducted revealing significant effects of the mode of administration. In particular, a significant main effect of mode of administration was found ($F_{1,48}$=33.91, $p<.01$) with students in the computer-administered condition (NO: $\bar{x} = 21.45$; CS: $\bar{x} = 14.82$) outperforming students in the paper-based administration (NO: $\bar{x} = 15.63$; CS: $\bar{x} = 9.15$). Correlational analyses corrected for attenuation further illustrated a lack of equivalency across modes as the NO subtest yielded strong cross-modal correlational values, ranging from .81 to .98, while the CS subtest yielded a wider range of values, .38 to .84. The reasons underlying the failure to find statistical equivalency between CS subtests presented in computer-based and paper-based modes are not well understood. At a minimum it is clear that the test-taking experience for students was dissimilar in the computer-based and paper-based format.

An examination of the aforementioned studies leads to the expectation that statistical equivalence between CBTs and P&Ps may only be established if the test taking

experience is paralleled across modes of administration; presentation and editorial control (i.e., the ability to revise, review, and skip questions) available to the examinee must be identical across forms. Furthermore, a standardized method or procedure must be established with which to confidently state that statistical equivalence has been achieved. Equivalency of CBTs and P&Ps must be resolved if CBTs are to effectively succeed or supplement the traditional P&P tests.

As directed by the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1986) "a clear rationale and supporting evidence should be provided for the claim that scores earned on different forms of a test may be used interchangeably" (p. 57). Further to this, in the APA's *Guidelines for Computer-Based Tests and Interpretations* (1986) it is stated that equivalence will be met when

> (a) the rank orders of scores of individuals tested in alternative modes closely approximate each other, and (b) the means, dispersions and shapes of the score distributions are approximately the same, or have been made approximately the same by rescaling the scores from the computer mode. (p. 18)

It is clear that simple correlational analyses are insufficient for determining the statistical equivalence of differing modes of administration; this insufficiency in the empirical literature illustrates the need for more rigorous methods or criteria for establishing cross-modal equivalence.

In an attempt to incorporate more rigorous methods for establishing cross-modal equivalence, a recent study by Wang and Kolen (2001) incorporated Lord's (1980) notion

of *equity* in examining equivalence results from CBTs and P&Ps. *First-order* equity is satisfied through equality of standardized scale scores (total test scores) across test forms, at all levels of examinee ability, going beyond simple correlation:

$$E(s_1 \mid \theta) = E(s_2 \mid \theta), \text{ for all } \theta \qquad (1)$$

where E refers to the expected value of scale scores over examinees of a given ability, θ, and s_1 is the scale score on Test 1 and s_2 is the scale score on Test 2 (Wang & Kolen, 2001, p.29). What equation (1) means is that at any given ability level, an individual should be able to earn the same scale score across forms. Further, satisfying the notion of *second-order equity* requires that test forms evidence equality of test score variance across test forms, again at every level of examinee ability:

$$\sigma(s_1 \mid \theta) = \sigma(s_2 \mid \theta), \text{ for all } \theta, \qquad (2)$$

where σ is the standard deviation of scale scores over examinees of a given ability, θ. The two notions of first- and second-order equity may be summarized by the criterion of *equal probabilities of achieving passing scores,*

$$P(s_1 \geq s_c \mid \theta) = P(s_2 \geq s_c \mid \theta), \text{ for all } \theta, \qquad (3)$$

where P is the probability of passing and s_c is the passing score. These criteria can be empirically assessed. First order equity is determined through calculation of the mean difference between mode scale scores:

$$Sdiff_\theta = \bar{s}_{\theta CAT} - \bar{s}_{\theta P\&P}. \qquad (4)$$

Hypothesized to be zero, $Sdiff_\theta$ is the difference in scale scores at any given ability, θ, and $\bar{s}_\theta$ is the mean scale score at θ for the mode of administration; second order equity is determined through calculation of the *conditional standard error of measurement of scale score* (*CSEMSS*), as given by the formula:

$$CSEMSS_\theta = \sqrt{[\Sigma(s_\theta - s_{(mean)\theta})^2]}. \quad (5)$$

Using the American ACT Mathematics exam, Wang and Kolen's (2001) equity analyses indicated that neither first- nor second-order equity were satisfied. In particular, response data from previously administered paper-and-pencil exams were used to calibrate a 420-item pool. Then, a CAT simulation system specifically programmed for the study generated 400 examinee responses at seventeen points on the theta scale for the computer-based administration. Paper-and-pencil responses were simulated according to the same procedure while keeping the items constant across replications. Based on this simulation, cross-modal comparisons showed non-equivalence between the two modes of administration. Differences in both true score and scale scores (*Sdiff*) were found as well as differences in true and scale score error variability (*CSEMSS*). Specifically, Wang and Kolen (2001) state that "71% of the examinees earn P&P number-correct based scale scores of 18 or higher. However, approximately 68% of the examinees earn CAT based scale scores above 18" (p. 45). The importance of such "inequity" becomes apparent when such a difference is seen to affect 204 examinees within this simulation. "A score of 18 on the ACT assessment is used as a cut score for collegiate sports eligibility by the NCAA" (Wang & Kolen, 2001, p. 45). While this example utilizes a computerized-adaptive test, it is nonetheless a good illustration of the increasingly rigorous statistical procedures required for assessing equivalence across modes of administration. The results presented show that statistical equivalence should not be assumed but thoroughly examined and established using increasingly rigorous methods.

Cognitive Comparability

Technological advancements continue to offer enormous potential in our ability to study and measure performance. Given the rapid rate of such developments, it is important that we understand technological instruments in the context of comparison with the standard, paper-and-pencil, instruments. Thorough examination of psychometric and cognitive issues associated with computerization will establish a theoretical and empirical foundation from which the full potential of this medium may be confidently used. Computerization offers great potential to test developers in the ability to present tasks that begin to approach real-life situations. Currently our understanding of performance on CBTs in comparison to traditional P&Ps is tenuous as evidenced by the failure to establish statistical equivalency between these two modes in many instances (e.g. Epstein & Klinkenberg, 2001). Through the application of cognitive psychological theory and analysis, performance within computer-based assessment situations may be better understood.

Cognition and assessment, while well established domains of research in their own right have only recently begun to merge (Embretson, 1999; Pellegrino, Baxter, & Glaser, 1999; Sternberg, 1984). Cognitive theory can be used to inform the development of test items as well as be used as a framework for the analysis and interpretation of test results. Convergence of these two fields can be seen in the investigation of processes and strategies evoked in the successful or unsuccessful completion of items, and in the use of those investigations to guide item development.

Previous research has failed to provide conclusive evidence regarding the comparability of computerized test forms with paper-and-pencil test forms (Divgi &

Stoloff, 1986; Fulcher, 1999; Greaud & Green 1986; Lee, Moreno, & Sympson, 1986; Reinking & Schreiner, 1985; Van de Vijver & Harsveld, 1994). Such discrepancies may result from the absence of both a standardized definition, and test of equivalency. Alternatively, such discrepancies may result from a lack of understanding of examinee performance within the environments of the two modes of test administration. As suggested by Spray et al. (1989) and Honaker (1988), the examinees' experience of the test-taking process can differentially affect performance if the administration procedures are not the same across computer- and paper-based modes; equivalence must be determined statistically and comparability must be established experientially, with respect to cognitive factors.

Cognitive factors hypothesized to affect comparability between CBTs and P&Ps have included the amount of time spent on the task or exam (Green, 1991), individually presented or group presented items (Dimock & Cormier, 1991), and the ability to review or skip items and return to them later (Vispoel, Rocklin, & Wang, 1994; Wise & Plake, 1989), just to name a few. Recent studies have begun to incorporate the above factors in the design of CBTs (Goldberg & Pedulla, 2002; Luecht et al., 1998; Mason, Patry, & Bernstein, 2001), meeting, however, with inconsistent conclusions as to the comparability of the two modes of administration.

In one study, Goldberg and Pedulla (2002) required two hundred and twenty-two students to respond to one of three conditions of the GRE General: a P&P condition, a computerized without editorial control condition, and a computerized with editorial control condition. In this instance, editorial control refers specifically to the ability to change a response after finalizing the answer as well as skipping an item to return to it

later. Administered under conditions parallel to actual GRE procedures, this study found a significant effect of mode of administration ($F_{6,412} = 5.443$, $p<.01$), across all three subtests (Analytical, Verbal, and Quantitative), with examinees in the P&P condition ($\bar{x} = 17.4, 24.1, 22.3$; $\sigma = 4.84, 5.54, 4.42$) outperforming those in the computerized without editorial control condition ($\bar{x} = 14.6, 21.6, 19.3$; $\sigma = 5.17, 5.91, 4.71$). Examinees in the P&P condition and those in the computerized with editorial control condition did not significantly differ in their performance. Consideration of cognitive factors as illustrated in the Goldberg and Pedulla (2002) study represents a step towards Bennett's (1998) *second-generation* CBTs, based on *cognitively principled design* (Bennett, 1999). However, it is through the application of cognitive psychological principles to user interface (UI) design that we may expect to see a greater possibility for establishing comparability and statistical equivalence (Booth, 1991, 1998; Huff & Sireci, 2001).

The UI is defined as the layer of the CBT that the examinee directly interacts with and it is this factor that acts as a direct moderator of computer-based performance (Booth, 1991). To ensure cognitive comparability, the UI design of CBTs must consider not only the presentation of the items but also the mental representation students generate of those items (in order to protect against construct-irrelevant variance) (AERA et al., 1986; APA, 1986; Huff & Sireci, 2001). Should it be found that the cognitive demands of CBTs are different from P&Ps, then comparability between the two modes is bound to be compromised. For example, if examinees are not permitted the use of an external memory aid in a CBT administration, it is possible that the UI may overload working memory, imposing complex processing demands "instead of [allowing students to use] their

limited information-processing resources to maintain [problem-based] information in working memory" (Pellegrino, 1995, p. 11).

Working memory has been defined as a cognitive system of finite resources in which information is temporarily stored and manipulated by three components: the phonological loop, the visuo-spatial sketchpad (VSSP), and the central executive (Baddeley, 1986, 2002; Baddeley & Hitch, 1974). The first component, the phonological loop, serves to store and manipulate verbal information as subvocal articulation; the VSSP serves to temporarily maintain and perform operations on visually and spatially presented information; and the central executive functioning serves to regulate the operations of the various other working memory sub-systems as well as moderating interaction with long-term memory (see Baddeley, 2002). Complex problem solving requires that information be manipulated in working memory; increasing demands on the subsystems of working memory can result in decreased task performance depending on the nature of the task.

Application of cognitive science research, specifically research on working memory, has the potential of informing researchers about the potential sources of statistical inequivalence and cognitive dissimilarities in testing. Working memory demands may differ across modes of administration, CBT and P&P, and as a result performance may differ yielding unequal assessment instruments.

There has been substantial research dedicated to investigating the effects of working memory demands on problem solving performance, specifically on the provision of aids to facilitate performance (e.g. Cary & Carlson, 1999, 2001; Hegarty & Steinhoff, 1997; Intons-Peterson, 1993; Intons-Peterson & Fournier, 1986; Schönpflug, 1986;

Trafton & Trickett, 2001). External memory aids are "any device or mechanism, external to the person, whose purpose is to facilitate memory" (Intons-Peterson, 1993, p. 135), often operationalized through the ability to take notes while performing a problem solving task (Cary & Carlson, 1999, 2001; Hegarty & Steinhoff, 1997; Trafton & Trickett, 2001). Cary and Carlson (1999, 2001) found that the ability to take notes throughout complex problem solving tasks – for example, calculating gross incomes based on a commission system (Cary & Carlson, 1999) and a complex arithmetic task (Cary & Carlson, 2001) - resulted in significant improvement in performance over those conditions in which participants were not given the option to take notes.

Hegarty and Steinhoff (1997) reinforced the role played by external memory aids in problem solving, directly linking performance to the VSSP sub-system of working memory in an experimental study of mechanical problem solving. Screening participants according to visuo-spatial ability using the Paper Folding Test (Ekstrom, French, Harman, & Dermen, 1976), these investigators found that low ability participants benefited significantly from the ability to take notes, thus relieving the load on the VSSP and showing improved performance (Hegarty & Steinhoff, 1997). Typically overlooked as a performance factor in CBTs, the ability to take notes during problem solving was examined by Trafton and Trickett (2001) where they incorporated a note pad into the UI of a computerized instrument. The importance of note taking in relieving working memory has been demonstrated in research on P&Ps, resulting in improved performance (Cary & Carlson, 1999, 2001; Hegarty & Steinhoff, 1997; Trafton & Trickett, 2001), but must now be addressed in CBT research in order to begin to understand sources of inequity.

Syllogistic Reasoning

Recently, educational measurement researchers have begun to pay attention to students' higher-order thinking ability (e.g., Embretson, 1999; Pellegrino, et al. 1999; Sternberg, 1984). Emphasis within the domain of measurement has shifted from pure content/domain assessment to performance assessment (Pellegrino et al., 1999). Within cognitive science, syllogistic reasoning in particular lends itself well to study of higher-order thinking because categorical syllogisms "lie on the borderline of human competence" (Johnson-Laird & Bara, 1984, p. 2). Categorical syllogisms range in difficulty – some syllogisms are very easy and answered correctly by most individuals, whereas others are very difficult and most individuals get them wrong. Johnson-Laird and Bara (1984) state that syllogisms are excellent measures of higher-order thinking ability as "they are at the center of quantificational reasoning; they have only a small number of different forms; and performance with them varies considerably…" (p.3).

Syllogisms present a set of two premises, each containing a single quantifier, that require the reasoner to deduce a conclusion that relates each of the end terms to one another (see Bucciarelli & Johnson-Laird, 1999; Johnson-Laird, 1983; Johnson-Laird & Bara, 1984 for a review of syllogistic reasoning). Syllogisms are often presented in a form similar to the following:

All beekeepers are painters	(Premise 1)
All painters are astronauts	(Premise 2)
All beekeepers are astronauts	(Conclusion)

The premises may, alternately, be presented in a more abstract form, such as "All A are B; All B are C." Four different quantifiers can also be used in each of the statements creating what are termed 'moods':

All A are B (mood A)

Some A are B (mood I)

No A are B (mood E)

Some A are not B (mood O)

Within the syllogism, the two premises may be presented as one of four 'figures':

A – B	B – A	A – B	B – A
B – C	C – B	C – B	B – C

Following the above premises, the conclusions may take the form of A – C or C – A, yielding a possible 64 forms within one figure. Syllogisms represent group inclusion problems of varying complexity; different combinations of premises require the generation and consideration of a variety of models or representations of the premises in order to yield correct conclusions.

Johnson-Laird and Bara (1984) explain the process associated with generating conclusions to syllogisms in terms of *mental models* (see Bucciarelli & Johnson-Laird, 1999; Johnson-Laird, 1983; Johnson-Laird & Bara, 1984 for a review of the theory of mental models). In using mental models, reasoners represent the information contained in the premises as mental tokens: "entities in a specific state of affairs" (Johnson-Laird & Bara, 1984, p. 4). The notation adopted by Johnson-Laird and Byrne (1991) and further clarified by Bucciarelli and Johnson-Laird (1999) proposed that the premise "All A are B," for example, can be represented using the following model:

[a] b

[a] b

where the two A's represent the set of A's in its entirety, indicated by the square brackets. Similarly, syllogisms of mood E (No A are B) can be represented as follows:

[a] -b

[a] -b

[b]

[b]

with the negative indicator (-) signifying no co-occurrence of A with B.

These mental models may take the form of vivid images or abstract visuo-spatial representations. However the models are represented, they are subject to the constraints imposed by the reasoner's working memory. The models of the premises must be generated and manipulated in order to deduce a necessary conclusion that relates the end terms found in the premises or to realize that a necessary conclusion is impossible.

Dependent on the mood and figure of the syllogism, specific premise combinations require the generation of one model or multiple models in arriving at a valid conclusion (Bucciarelli & Johnson-Laird, 1999). In contrast, other premise combinations result in no valid conclusion (NVC). Examples of each type of syllogism and the corresponding mental model notations are as follows:

One-model									
All A are B	[a]	[b]	c						
All B are C	[a]	[b]	c						
∴ All A are C			c						
Multiple-model									
Some A are B	a	[b]	-c	a		[c]	a		[c]
No B are C	a			a	[b]	-c	a	[b]	-c
∴ Some A are not C		[b]	-c		[b]	-c		[b]	-c
			[c]			[c]	a		[c]
						[c]			
NVC									
All A are B	[a]	b	-c	[a]	b	c			
Some B are not C	[a]	b	-c	[a]	b	c			
∴ NVC					b	-c			
					b	-c			

Syllogisms requiring the creation and manipulation of multiple models in arriving at a conclusion are more difficult than those requiring fewer models (Bucciarelli, & Johnson-Laird, 1999; Johnson-Laird & Bara, 1984; Johnson-Laird & Byrne, 1991). For example, Bucciarelli and Johnson-Laird (1999) conducted two experiments with twenty undergraduate psychology volunteers. Volunteers wrote their conclusions to twenty syllogisms (4 one-model, 8 multiple-model, and 8 NVC syllogisms); both experiments were designed similarly with the second experiment recording "think aloud" data from the participants. Both of the experiments yielded data that very clearly illustrate the effect that the number of models has on performance. One-model problems were solved correctly by 80% of participants in experiment 1 and 76% of participants in experiment 2. In contrast, multiple-model syllogisms were markedly more difficult as only 21% of participants correctly solved the syllogisms across both experiments. Furthermore, no valid conclusion problems were moderately difficult; fifty percent of participants in

experiment 1 solved the syllogisms correctly and only 41% of participants in experiment 2 solved the syllogisms correctly.

One proposed explanation for such differences is that syllogisms requiring the manipulation of a greater number of models tax a limited-capacity working memory – previous studies have shown that loading working memory decreases performance on syllogistic reasoning tasks (Fisher, 1981; Gilhooly, Logie, Wetherick, & Wynn, 1993; Gilhooly, Logie, & Wynn, 1999, 2002). Reducing the load on working memory through such external memory aids as note pads should serve to facilitate syllogistic reasoning performance, as has been shown in other problem solving studies (Cary & Carlson, 1999, 2001; Trafton & Trickett, 2001).

Paper Folding Test

The Paper Folding Test is a standardized test provided by Educational Testing Services as part of the *Kit of Factor-Referenced Cognitive Tests* (Ekstrom et al., 1976). The purpose of this test is to provide a measure of *visualization*, "the ability to manipulate or transform the image of spatial patterns into other visual arrangements" (Ekstrom et al., 1976, p. 41). The test is comprised of two sections of ten items each requiring the examinees to attend to a figure of a folded sheet of paper with holes punched in it and then to select from a list of alternatives the figure representing the way in which the folded sheet would appear if unfolded (see Appendix B). Scores on this test provide a measure of visuo-spatial ability or capacity, with high scores indicating that an examinee can easily store and manipulate information in the visuo-spatial sketchpad (VSSP) whereas low scores indicate difficulty with visuo-spatial information. This test was used by Hegarty and Steinhoff (1997) to sort examinees according to their visuo-

spatial ability in an effort to explain performance on a complex problem solving task with or without provision of an external memory aid.

Current Study

The current study will examine the equivalency and comparability of CBTs and P&Ps regarding higher-order thinking within the context of two memory aid conditions: use of a notepad or no use of a notepad. Equivalency of CBTs and P&Ps has yet to be established in the psychometric literature (Epstein & Klinkenberg, 2001). Non-standardized statistical determinants of equivalency and a lack of research studies focusing on cognitive factors influencing comparability have failed to resolve the issue of equivalence and comparability between CBTs and P&Ps. Paper-and-pencil assessments provide examinees with the opportunity to take notes on the instrument itself, serving as a memory aid, thereby decreasing the load on working memory and, in particular, the visuo-spatial sketchpad (Hegarty & Steinhoff, 1997). Recent shifts to CBTs have not incorporated this feature; examinees simply select the desired response in the case of multiple-choice exams without the advantage of a note pad or similar aid. Guided by Lord's notion of *equity* (1980) as implemented by Wang and Kolen (2001), differential item functioning (DIF) will be assessed across modes of administration and memory aid conditions. DIF is defined as the probability that examinees at similar observed scores will select distinct response options.

Research Questions

1) How is students' complex problem solving (syllogistic) performance affected by computer- versus paper-based modes of test administration?
2) How does loading of working memory (WM) affect syllogistic performance?
    a. Among those students allowed to use a memory aid, does actual use of an external memory aid result in better performance?
    b. Given that students possess low or high VSSP ability, how is performance affected provided an external memory aid?
3) How do factors associated with WM interact with mode of administration?
4) Are there any differences among groups across mode of administration and memory load at either the test or item level?

CHAPTER II - METHODS

Participants

Two hundred and sixteen undergraduate students from two introductory psychology courses at the University of Alberta served as volunteers for the present study. The students are required to complete two or three hours of research participation in partial fulfillment of course credit. Students are recruited through a research participation web site.

Design

A 2 X 2 fully crossed between-subject factorial design was used in the present study. The first factor involved mode of administration (computer-based administration versus paper-based administration). Each participant was randomly assigned to complete the task in either the computer- or paper-based environment. The second factor involved availability of a memory aid (sheets of paper versus no sheets of paper). One-half of participants were randomly assigned to the *memory-aid* level of this factor, in which they were encouraged to take notes on the provided notepad while performing the syllogistic reasoning task. The other half of the participants was assigned to the *no-memory-aid* condition, in which they were prohibited from taking any notes during the task. The dependent variable in the present study is accuracy on the syllogistic reasoning item set.

Procedure and Materials

Participating students were tested in groups of approximately ten in a classroom. Once consent was obtained from each of the participants, those in the computer-based

condition were presented with the following tasks on the computer screen: ten demographic questions, twenty-four syllogisms, and the Paper Folding Test (see Appendix C). The tasks were programmed in Hyper Text Markup Language (HTML) as a web page. Participants responded using the mouse to click on the desired options. Each participant was seated at a computer terminal that had the study web page preloaded and presented on the screen. The tasks were presented as follows: page one contained the demographic items; page two introduced the syllogistic reasoning task; pages three and four presented twelve syllogisms each; page five provided instructions for the paper folding test; pages six and seven presented ten paper folding items each; page eight concluded the study, thanking the participants and providing the study debriefing. The tasks were presented within an Internet Explorer window of fixed dimensions; if the page currently loaded required more space then immediately available within the window, participants were able to scroll through the whole page. Only a single response to each option could be made and this response could be changed any number of times prior to moving from one page to the next. Editorial control available to participants was similar, however not exactly the same as that available within paper-based condition. Participants could change or skip items and change responses within each page. Once a page was completed it was submitted and could not be returned to. Response data was emailed to the researcher at the completion of each page. Participants in the paper-based condition were presented with the same tasks but in a paper-based format. As with the computer-based administration, participants in the paper-based condition completed the tasks in the same order and with the same layout (e.g. page 1 contained demographic items, etc.).

The twenty-four syllogisms were presented in one of ten randomized orders, while maintaining the standard order of the demographic and Paper Folding Test items. The categories that made up the topics of the syllogisms involved occupations described by single words, selected from the National Occupation Classification List (Citizenship and Immigration Canada, n.d., Table 1). The twenty-four syllogisms presented to participants included 8 one-model syllogisms, 8 multiple-model syllogisms, and 8 no valid conclusion (NVC) syllogisms. In addition, each of the four moods was represented by 6 syllogisms. For each syllogism, a list of five conclusions was presented to participants. For example, the conclusions presented for the syllogism:

All Painters are Politicians

All Politicians are Bankers

included: All Painters are Bankers; Some Painters are Bankers; No Painters are Bankers; Some Painters are not Bankers; Nothing. Keyed options were selected based on their logical necessity given the premise set as reported by Johnson-Laird and Bara (1984).

For participants assigned to the memory aid condition, two sheets of blank paper were provided. A maximum of one hour was provided for the participants to complete the study. In the memory-aid condition, participants were encouraged to use only the blank paper provided to help them solve the syllogisms. Participants were told that no other external aid could be used to solve the problems. If requested, additional paper was provided to the participants. For the no-memory-aid condition, participants were instructed to not use anything to help them complete the tasks.

Hypotheses

Hypothesis 1: Mode of Administration

A: Item p-values will not differ as a result of mode of administration.

$_{1a}h_0$: $r_{CBT*P\&P} = 1.0$ $_{1a}h_1$: $r_{CBT*P\&P} < 1.0$

B: Syllogistic performance will not differ as a result of mode of administration.

$_{1b}h_0$: $\mu_{CBT} = \mu_{P\&P}$ $_{1b}h_1$: $\mu_{CBT} \neq \mu_{P\&P}$

C: Syllogistic model types will not evidence DIF as a result of mode of administration.

$_{1c}h_0$: $\beta_F = 0$ $_{1c}h_1$: $\beta_F > 0$

Hypothesis 2: External Memory Aid

A: Item p-values will not differ as a result of provision of a memory aid.

$_{2a}h_0$: $r_{AID*NO\text{-}AID} = 1.0$ $_{2a}h_1$: $r_{AID*NO\text{-}AID} < 1.0$

B1: Syllogistic performance will not differ as a result of provision of a memory aid.

$_{2b1}h_0$: $\mu_{AID} = \mu_{NO\text{-}AID}$ $_{2b1}h_1$: $\mu_{AID} \neq \mu_{NO\text{-}AID}$

B2: Provided a memory aid, syllogistic performance will not differ as a result of use of the memory aid.

$_{2b2}h_0$: $\mu_{AID*USE} = \mu_{AID*NO\text{-}USE}$ $_{2b2}h_1$: $\mu_{AID*USE} = \mu_{AID*NO\text{-}USE}$

C: Syllogistic items will not evidence DIF as a result of provision of a memory aid.

$_{2c}h_0$: $\beta_F = 0$ $_{2c}h_1$: $\beta_F > 0$

Hypothesis 3: Interaction

A: There will be no interaction of mode of administration and memory aid provision.

$_{3a}h_0$: $(\alpha\beta)_{11} = (\alpha\beta)_{12} = (\alpha\beta)_{21} = (\alpha\beta)_{22}$

α_1 = CBT α_2 = P&P

β_1 = Aid β_2 = No Aid

$_{3a}h_1$: At least one cell mean under h_0 does not equal all others.

B: Provided a memory aid, there will be no interaction of mode of administration and memory aid usage.

$_{3b}h_0$: $(\alpha\gamma)_{11} = (\alpha\gamma)_{12} = (\alpha\gamma)_{21} = (\alpha\gamma)_{22}$

α_1 = CBT α_2 = P&P

γ_1 = Use γ_2 = No Use

$_{3b}h_1$: At least one cell mean under h_0 does not equal all others.

C: Syllogistic items will not evidence DIF as a result of interaction of mode of administration, memory aid provision, and memory aid use:

$_{3c}h_0$: $\beta_R = 0$ $_{3c}h_1$: $\beta_R > 0$

Hypothesis 4: VSSP

$_4h_0$: Blocking on VSSP ability (low/high), the above hypotheses are true.

$_4h_1$: Blocking on VSSP ability (low/high), any of the above hypotheses are not true.

CHAPTER III - RESULTS

Demographics

Of the 216 participants in this study, responses from only 209 participants are included as 7 participants failed to comply with the instructions regarding the paper-folding task; these participants used the scrap paper provided to solve the task. The remaining participants in this study were 141 females and 68 males, a distribution expected of introductory psychology courses at the University of Alberta. The majority of the participants were 18-20 years of age (n = 167) and first year students (n = 116). Participants represented both faculties through which psychology courses are offered at the University of Alberta: Arts (n = 83) and Sciences (n = 101). A more detailed examination of the sample for this study can be seen in Table 2.

Across conditions, participants were asked to rate their ability in particular computer related tasks as well as their overall level of comfort in using computers. Self-ratings of skill with various computer-related tasks were also assessed on a 5-point likert scale, 1 = *none*, 2 = *poor*, 3 = *fair*, 4 = *good*, and 5 = *excellent*. As can be seen in Table 3, computer skills were rated as follows: word processing ($\bar{X} = 4.17$, $\hat{\sigma}_X = 0.73$), e-mail ($\bar{X} = 4.35$, $\hat{\sigma}_X = 0.69$), games ($\bar{X} = 3.40$, $\hat{\sigma}_X = 1.09$), and searching the web ($\bar{X} = 3.91$, $\hat{\sigma}_X = 0.74$). Participants' comfort with computers was assessed on a 5-point likert scale on which they would indicate their comfort, 1 = *not at all*, 2 = *slightly*, 3 = *somewhat*, 4 = *comfortable*, and 5 = *very*. Participants generally rated themselves as *comfortable* to *very comfortable* (n = 169) with no participants rating their comfort with computers as *not at all*.

Self-reported fluency in spoken English was also surveyed. As only two participants rated themselves as "non-fluent" this variable was not included in subsequent analyses.

Participants were randomly assigned to one of the four experimental conditions: paper*aid (n = 52), paper*no aid (n = 58), computer*aid (n = 49), and computer*no aid (n = 50).

Item Correlation

The results of this study evidenced strong correlations of syllogistic performance across conditions through examination of item p-values. P-values indicate item difficulty, reporting the proportion of participants who correctly answered each item (Table 4). Across mode of administration, paper or computer, item p-values were strongly correlated, $r = 0.97$, $p<.05$. Item p-values were also strongly correlated between memory aid conditions, $r = 0.97$, $p<.05$ (Table 5). The syllogistic instrument itself displayed strong internal consistency, $\alpha = 0.81$.

Test-Level Analyses

A 2 X 2 ANOVA comparing syllogistic performance across mode of administration (computer or paper) and memory aid provision (aid or no aid) was conducted (Table 6). As shown in Table 7, a main effect of mode of administration was found, $F_{1,205} = 4.58$, $p<.05$, with participants in the paper condition, $\bar{X} = 14.79$, $\hat{\sigma}_X = 4.49$, scoring significantly higher than participants in the computer-based condition, $\bar{X} = 13.51$, $\hat{\sigma}_X = 4.13$ (see Table 6). No significant main effect of memory aid was found nor was a significant interaction of memory aid and mode of administration found.

Within the memory aid condition, those participants who actually used the memory aid were identified from those who did not. After identification, another ANOVA was conducted to determine if participants who used a memory aid differed in their syllogistic performance from those who did not use it, across modes of administration (Tables 8 and 9). A significant main effect of memory aid use was found, $F_{1,97} = 13.19$, $p<.05$, with participants who used the memory aid performing better, $\bar{X} = 16.11$, $\hat{\sigma}_X = 4.45$, than those who did not use the memory aid, $\bar{X} = 13.19$, $\hat{\sigma}_X = 3.93$ (see Table 8). Further, the main effect of mode of administration approached significance, $F_{1,97} = 3.24$, $p<.10$. As above, participants in the paper-based condition performed better, $\bar{X} = 14.87$, $\hat{\sigma}_X = 4.68$, than participants in the computer-based condition, $\bar{X} = 13.67$, $\hat{\sigma}_X = 3.93$. There was no significant interaction of mode of administration and memory aid use.

The above analyses of variance were repeated, blocking on visuo-spatial ability as measured by the Paper Folding Test[1]. Visuo-spatial ability scores were calculated according to the following formula correcting for guessing provided in the *Kit of Factor-Referenced Cognitive Tests* (Ekstrom et al., 1976, p. 10):

$$R - W / n - 1, \qquad (6)$$

where R is the participant's number correct, W is the number incorrect, and n is the number of response options for each item. Participants were then categorized as high- or low-ability based on a median split, set at a score of 15.2 out of a maximum score of 20[2] (Hegarty & Steinhoff, 1997).

The sample was split into students of high visuo-spatial ability and low visuo-spatial ability (Tables 11 and 12). After splitting, a 2 X 2 ANOVA (mode of

administration X memory aid) was conducted for each separate sample of students. No main effects of mode of administration or memory aid were found for either sample. In addition, no interactions were found (Table 12).

Syllogistic performance was then more closely examined among participants who were provided a memory aid and actually used it.

[INSERT FIGURE 1 HERE]

For low-ability participants, indicated by the lines with the square end-points, a significant main effect of memory aid use was found, $F1,53 = 14.22$, $p<.05$, as participants who used the memory aid (solid line) performed significantly better, = 15.89, = 4.74, than participants who did not use the memory aid (dashed line), = 11.82, = 3.36 (see Table 13). No other main effects or interactions were discovered for either of the low or high ability groups (Table 14).

Item-Level Analyses

Item level analyses were conducted using the TESTGRAF software package (Ramsay, 2000). TESTGRAF uses nonparametric regression with kernel smoothing (Ramsay, 1991) to statistically analyze the differential item functioning (DIF). This method of analysis estimates and displays the relation between examinees at different observed scores and the probability that they will select the correct multiple-choice option for each item. This relationship is graphically described through an item characteristic curve (ICC).

[INSERT FIGURE 2 HERE]

Nonparametric regression can also compare differences across two or more groups of examinees, matching them on overall test performance. Denoted as β, DIF is

summarized by quantification of the area between the item characteristic curves. In comparing two groups, the group thought to be disadvantaged is labelled the focal group (F) and the group to which it is compared is the reference group (R). The index resulting from the comparison of the two groups, β_F, is the weighted score difference between the item characteristic curves for examinees of the same proficiency across the two groups. A negative β_F indicates that the reference group is advantaged while a positive β_F indicates that the focal group is advantaged. Comparing multiple groups, a composite DIF index is also computed (β_R), indicating the total DIF across all focal groups in comparison to a single reference group. This value is always positive and is the root mean square average difference for each item, between each focal group and the reference group.

While there is no statistical test for TESTGRAF, guidelines have been previously established for interpreting the DIF index (Gierl & Bolt, 2001; Nandakumar, 1993). A-level DIF, $|\beta_F| < 0.05$, is considered negligible; B-level DIF, $0.05 \leq |\beta_F| < 0.10$, is considered moderate; and C-level DIF, $|\beta_F| \geq 0.10$, is considered large. Typically, B- and C-level DIF call for closer examination of the item(s) in question. Guidelines for interpreting β_R have yet to be established, however, given that β_R is the average difference for each item, the single-item guidelines are often used for interpreting composite DIF.

DIF analysis in syllogistic performance data was conducted using TESTGRAF, comparing across groups the probability of correctly answering each item given a specific total score (see Appendix D). For this study, the groups are the defined by the experimental variables: paper-based administration with memory aid, paper-based administration with no memory aid, computer-based administration with memory aid,

computer-based administration with no memory aid. Analyses were further conducted within the memory aid condition by comparing use or non-use of the memory aid in both the paper- and computer-based administration conditions. As items of similar type are expected to function similarly and also given that many items evidence $\beta_F > .05$, items were grouped according to syllogism model type – one-model, multiple-model, or no valid conclusion (NVC) – and examined according to the presence and direction of C-level DIF. For each item type, DIF may be present favouring both the reference and focal groups as DIF is said to be present in this study if at least one of eight items of each of the syllogistic model types evidences C-level DIF. DIF analyses were also conducted using the full sample of examinees as well as using sub-samples, split according to visuo-spatial ability (see Tables 15 – 20 for beta values from each DIF analysis). The above DIF analyses are summarized and presented graphically, as in the figure below:

[INSERT FIGURE 3 HERE]

Each graph represents a syllogistic item type within the overall ability, low ability, and high ability groups across the full experimental conditions or within the memory aid provided condition. The eight syllogisms of the type represented – one model, multiple-model, or NVC – are displayed as points within columns representing each of the experimental comparisons, labelled according to the focal group. C-level DIF, $|\beta_F| \geq 0.10$, is indicated by a dashed line on each graph. Items falling within the band of dashed lines indicate A- or B-level DIF, whereas items falling beyond the band (either above or below) indicate C-level DIF.

The first DIF analysis involved comparing the paper-based condition (the reference group) to the computer-based condition (the focal group) for the full sample.

This analysis was motivated by an interest to see if examinees' performance on syllogisms favoured the paper- or the computer-based condition. The results of this analysis can be seen in the first column of Figures 4, 5, and 6. Very little DIF is evident, although some DIF is evident within the one-model syllogisms favouring the focal group, computer-based administration. The second DIF analysis involved comparing the memory aid condition (the reference group) to the no memory aid condition (the focal group) (see the second column of Figures 4, 5, and 6). Again, this analysis showed very little DIF. None of the items favoured the reference group or the focal group. The third DIF analysis involved multiple comparisons between the paper-based and memory aid condition (i.e., paper*aid; the reference group) and the following set of conditions: paper*no-aid, computer*aid, computer*no-aid (see the third, fourth, and fifth columns of Figures 4, 5, and 6). The latter conditions were each set as focal groups in the analysis. Results of this analysis revealed DIF favouring the paper*aid condition over the following conditions: computer*aid for multiple-model syllogisms; computer*no-aid for multiple-model syllogisms; and paper*no-aid conditions for one-model, multiple-model, and NVC syllogisms. In contrast, DIF was evidenced favouring the computer*aid condition (focal group) over the paper*aid condition (reference group) for one-model syllogisms. In addition, DIF was evidenced favouring the computer*no-aid conditions (focal group) over the paper*aid condition (reference group) for one-model and NVC syllogisms.

The above DIF analysis procedure was then repeated examining memory aid use or non-use within the *memory aid* condition. This DIF analysis first compared paper- (reference group) and computer-based administration (focal group) given that a memory

aid was available (see the first column of Figures 7, 8, and 9). This analysis showed DIF favouring paper-based administration for multiple-model and NVC syllogisms as well as DIF favouring computer-based administration for one-model and NVC syllogisms. DIF analysis was then used to compare actual use of the memory aid (the reference group) to non-use of the memory aid when available (the focal group). DIF was evidenced favouring use of the memory aid for multiple-model and NVC syllogisms as well as favouring non-use of the memory aid when available for one-model and multiple-model syllogisms (see the second column of Figures 7, 8, and 9). As before, a third DIF analysis involved multiple comparisons between the paper-based administration with use of memory aid (i.e., paper*use; the reference group) and the following: paper*non-use, computer*use, and computer*non-use groups (see columns three, four, and five of Figures 7, 8, and 9). Results of this analysis revealed DIF favouring the paper*use group (reference group) over the following groups: computer*use (focal group) for multiple-model and NVC syllogisms; computer*non-use for multiple-model and NVC syllogisms; and paper*non-use for multiple-model and NVC syllogisms. In contrast DIF was evidenced favouring the focal group over the paper*use group (reference group) for each of the following groups according to the multiple comparison DIF analyses: computer*non-use for one-model and multiple-model syllogisms; paper*non-use for one-model and multiple-model syllogisms.

The above DIF analysis procedures were repeated for the sub-groups of low- and high visuo-spatial ability examinees.

The following are the DIF analyses for the low visuo-spatial ability group. The first DIF analysis involved comparing the paper-based condition (the reference group) to

the computer-based condition (the focal group). The results of this analysis can be seen in the first column of Figures 10, 11, and 12. Very little DIF is evident as the multiple-model syllogisms favour the reference group, paper-based administration. The second DIF analysis involved comparing the memory aid condition (the reference group) to the no memory aid condition (the focal group) (see the second column of Figures 10, 11, and 12). DIF is evident favouring the aid condition (the reference group) for one-model and multiple-model syllogisms. DIF was also present favouring the no-aid condition (the focal group) for NVC syllogisms. The third DIF analysis involved multiple comparisons between the paper-based and memory aid condition (i.e., paper*aid; the reference group) and the following set of conditions: paper*no-aid, computer*aid, computer*no-aid (see the third, fourth, and fifth columns of Figures 10, 11, and 12). The latter conditions were each set as focal groups in the analysis. Results of this analysis revealed DIF favouring the paper*aid condition over the following conditions: computer*aid for multiple-model and NVC syllogisms; computer*no-aid for one-model and multiple-model syllogisms; and paper*no-aid for one-model and multiple-model syllogisms. In contrast, DIF was evidenced favouring the computer*aid condition (focal group) over the paper*aid condition (reference group) for one-model and multiple-model syllogisms; favouring the computer*no-aid conditions (focal group) over the paper*aid condition (reference group) for one-model and NVC syllogisms; and favouring the paper*no-aid condition over the paper*aid condition for multiple-model and NVC syllogisms.

The above DIF analysis procedure was then repeated examining memory aid use or non-use within the *memory aid* condition. This DIF analysis first compared paper- (reference group) and computer-based administration (focal group) given that a memory

aid was available (see the first column of Figures 13, 14, and 15). This analysis showed DIF favouring paper-based administration for multiple-model and NVC syllogisms as well as DIF favouring computer-based administration for one-model and multiple-model syllogisms. DIF analysis was then used to compare actual use of the memory aid (the reference group) to non-use of the memory aid when available (the focal group). DIF was evidenced favouring use of the memory aid for multiple-model and NVC syllogisms as well as favouring non-use of the memory aid when available for one-model and multiple-model syllogisms (see the second column of Figures 13, 14, and 15). As before, a third DIF analysis involved multiple comparisons between the paper-based administration with use of memory aid (i.e., paper*use; the reference group) and the following: paper*non-use, computer*use, and computer*non-use groups (see the third, fourth and fifth columns of Figures 13, 14, and 15). Results of this analysis revealed DIF favouring the paper*use group (reference group) over the following groups: computer*use (focal group) for NVC syllogisms; computer*non-use for one-model, multiple-model, and NVC syllogisms; and paper*non-use for multiple-model and NVC syllogisms. In contrast DIF was evidenced favouring the focal group over the paper*use group (reference group) for each of the following groups according to the multiple comparison DIF analyses: computer*use for multiple-model syllogisms; computer*non-use for one-model and multiple-model syllogisms; and paper*non-use for one-model and multiple-model syllogisms.

Finally, the following DIF analyses are for the high visuo-spatial ability group. The first DIF analysis involved comparing the paper-based condition (the reference group) to the computer-based condition (the focal group). The results of this analysis can be seen in the first column of Figures 16, 17, and 18. DIF is evident as the reference

group, paper-based administration, was favoured for one-model and multiple-model syllogisms. Also, the focal group, computer-based administration, was favoured for one-model and NVC syllogisms. The second DIF analysis involved comparing the memory aid condition (the reference group) to the no memory aid condition (the focal group) (see the second column of Figures 16, 17, and 18). DIF is evident favouring the aid condition (the reference group) for NVC syllogisms. DIF was also present favouring the no-aid condition (the focal group) for one-model and multiple-model syllogisms. The third DIF analysis involved multiple comparisons between the paper-based and memory aid condition (i.e., paper*aid; the reference group) and the following set of conditions: paper*no-aid, computer*aid, computer*no-aid (see columns three, four, and five of Figures 16, 17, and 18). The latter conditions were each set as focal groups in the analysis. Results of this analysis revealed DIF favouring the paper*aid condition over the following conditions: computer*aid for multiple-model and NVC syllogisms; computer*no-aid for one-model and NVC syllogisms; and paper*no-aid for multiple-model and NVC syllogisms. In contrast, DIF was evidenced favouring the computer*aid condition (focal group) over the paper*aid condition (reference group) for one-model, multiple-model, and NVC syllogisms; favouring the computer*no-aid conditions (focal group) over the paper*aid condition (reference group) for one-model and multiple-model syllogisms; and favouring the paper*no-aid condition over the paper*aid condition for one-model and NVC syllogisms.

The above DIF analysis procedure was then repeated examining memory aid use or non-use within the *memory aid* condition. This DIF analysis first compared paper- (reference group) and computer-based administration (focal group) given that a memory

aid was available (see the first column of Figures 19, 20, and 21). This analysis showed DIF favouring paper-based administration for multiple-model and NVC syllogisms as well as DIF favouring computer-based administration for one-model, multiple-model, and NVC syllogisms. DIF analysis was then used to compare actual use of the memory aid (the reference group) to non-use of the memory aid when available (the focal group). DIF was evidenced favouring use of the memory aid for multiple-model and NVC syllogisms as well as favouring non-use of the memory aid when available for one-model, multiple-model, and NVC syllogisms (see the second column of Figures 19, 20, and 21). A third DIF analysis involved multiple comparisons between the paper-based administration with use of memory aid (i.e., paper*use; the reference group) and the following: paper*non-use, computer*use, and computer*non-use groups (see columns three, four, and five of Figures 19, 20, and 21). Results of this analysis revealed DIF favouring the paper*use group (reference group) over the following groups: computer*use (focal group) for one-model, multiple-model, and NVC syllogisms; computer*non-use for multiple-model and NVC syllogisms; and paper*non-use for one-model, multiple-model ,and NVC syllogisms. In contrast DIF was evidenced favouring the focal group over the paper*use group (reference group) for each of the following groups according to the multiple comparison DIF analyses: computer*use for one-model, multiple-model, and NVC syllogisms; computer*non-use for one-model and NVC syllogisms; and paper*non-use for one-model, multiple-model, and NVC syllogisms.

The DIF described above is also presented in Table 21, indicating presence and direction of C-level differential item functioning across ability, experimental conditions, and syllogistic model types. A discernable pattern in differences in item functioning can

be seen across model types and ability. Across conditions and ability levels, performance on one-model syllogisms most often favours the focal group, neither paper nor memory aid testing condition, while multiple-model and NVC-type syllogisms increasingly favour the reference group, a paper- or aid-based testing condition. As well, comparison of DIF across low- and high-ability examinees indicates definite shifts in performance. Low ability participants, as with the full sample, typically perform better within the context of the focal group conditions for one-model syllogisms; DIF is split between reference and focal groups for multiple-model syllogisms; NVC syllogistic performance is facilitated by paper-based administration and the availability of a memory aid. Within the low ability group, the above effects are notable for all syllogistic model types but most pronounced for one-model and, especially, NVC syllogisms. While DIF for the most difficult items, multiple-model syllogisms, is similar across abilities, high-ability participants performance on one-model and NVC syllogisms is greatly facilitated within the context of computer-based test administration without the benefit of an external memory aid. This evidences a reversal effect of DIF seen in the low-ability sample.

CHAPTER IV – CONCLUSION

Summary

This study examined the interaction of mode of test administration (paper-based versus computer-based) with external memory aid (the ability to take notes versus no note taking). As has been shown in previous studies of statistical equivalence between paper and computer-administered tests, this study evidenced a high correlation ($r = 0.97$) between modes of administration as well as the across memory aid provision and prohibition conditions ($r = 0.97$). Such a strong correlation indicates that items administered in one condition (i.e., paper or aid) are of the same, or similar, difficulty when administered in the opposite condition (i.e., computer or no aid). According to previous assumptions regarding statistical equivalence, this measure would indicate that the two modes of administration are interchangeable or comparable. Correlation, however, is simply a measure of rank ordering; in this case indicating that syllogistic item difficulty was very similar across conditions. Further analyses and examination of the testing environment, or experience, illuminate differences across P&Ps and CBTs suggesting that test-taking experiences across each of these modes are, in fact, not comparable.

In an attempt to adequately establish statistical equivalency or test comparability, in-depth analyses were conducted at both the test- and item- level, similar to those methods used by Wang and Kolen (2001). At the test level, experimental manipulation of the mode of administration of twenty-four syllogisms showed more than a one-point (5%) difference in performance, favouring the paper-based condition. In addition, although experimental manipulation of memory aid showed to not facilitate performance, those

students who actually used the aid exhibited better performance than those who did not use the aid.

At the item level, differential item functioning (DIF) of syllogistic item types across experimental conditions was found. Overall item performance consistently favoured the paper-based condition and the memory aid condition for the most difficult items – that is, the multiple-model syllogisms. This finding is consistent with theory and previous studies as P&P scores are typically reported to be higher than CBT scores (i.e., Dimock & Cormier, 1991; Divgi & Stoloff, 1986; Lee et al., 1986; Noyes & Garland, 2003; Van de Vijver & Harsveld, 1994). In addition, previous studies have also shown that a memory aid provides the opportunity to relieve WM, thus facilitating performance (Cary & Carlson, 1999, 2001; Hegarty & Steinhoff, 1997; Trafton & Trickett, 2001). This effect was also evident for NVC syllogistic item performance, typically favouring P&P and memory aid availability, however, not as strongly as in the case of multiple-model syllogisms. An interesting effect was also observed for the easiest items, one-model syllogisms. Unlike the two other, more difficult, item types, examinees generally performed better on one-model syllogisms in computer-based administrations and those conditions in which a memory aid was not available. The above findings can be explained in terms of mental models and WM capacity. In conditions that relieve WM by allowing students to externalize their representations, difficult items are answered correctly more frequently than in conditions that do not relieve WM.

Performance differences due to mode of administration and availability of a memory aid became more apparent when examinees were compared according to low or high visuo-spatial ability. As also shown by Hegarty and Steinhoff (1997), low ability

examinees in this study benefited most from the ability to take notes in a paper-based testing condition, as the use of an external memory aid relieved a limited capacity WM. The use of an external memory aid by low ability examinees in the current study showed more than a 12% performance increase over those examinees who did not use the aid (total score = 15.89 and 11.82, respectively). Syllogistic performance by low ability examinees who used the memory aid was similar to the performance of all high ability examinees (15.73) as well as being similar to the performance of examinees in the P&P condition who did not use the memory aid (15.94). Further, this effect was also seen to interact with mode of administration as low ability examinees that took advantage of the external memory aid in the P&P condition outperformed those in the CBT condition by 5%. This effect may be the result of the ease with which examinees interacted with the test interface – paper- or computer-based test administration.

The ability to relieve WM, whether through availability of an external memory aid or a paper-based testing condition, resulted in poorer performance for high-ability examinees, especially with regards to easier items (one-model syllogisms); a phenomenon that has been referred to as the Expertise Reversal Effect (Chalmers, 2003; Kalyuga, Ayres, Chandler, & Sweller, 2003). Thus far illustrated in reading comprehension and vocabulary learning (Kalyuga et al., 2003; Yeung, 1999), the Expertise Reversal Effect occurs when high-ability examinees, capable of effectively storing and manipulating information in WM, are presented with a redundancy of information or problem space. With an increased amount of information, the possible trajectories through the problem space also increases, making the problem difficult for examinees who are able to consider a larger number of solutions. High-ability examinees

in the present study could have been susceptible to the redundancy of information offered in the form of the note pad. The presence of a note pad could have caused high ability examinees to consider more possible solutions, thereby increasing the possibility of a mistake. Also, previous research (Noyes & Garland, 2003) has shown that computer-based test performance allows examinees to respond very simply to cognitive tasks, removing any need for consideration of manner of response given that very few actions are allowed within a computerized test interface. Redundancy of the problem space, therefore, could also come as a result of the easily manipulated paper-based test interface.

At this point it would be premature to say that P&Ps and CBTs may be considered to be equivalent, or even comparable, test forms. It would also be inadequate to simply state that paper-based administration of a syllogistic reasoning task results in higher performance when compared to computer-based administration

While syllogistic performance within this study was highly correlated across conditions, examinees' test experience and working memory ability are shown to be contributing factors in establishing that CBTs and P&Ps are not comparable. Accounting for a simple feature within the testing experience, the ability to take notes, served to illuminate performance differences contributing to differential test functioning. Test performance is maximized when memory loading, the cognitive counterpart of item difficulty, is incorporated in test interface design. Difficult items are more likely to be solved correctly if the load on working memory is relieved; however, these effects reverse as the visuo-spatial ability of the examinee increases.

Discussion

The question is no longer 'Does performance differ across modes?' but 'How does performance differ across modes?' For answers we turn our attention to the domain of cognitive psychology, with specific attention to issues associated with test and interface design (i.e., Booth, 1991, 1998; Huff & Sireci, 2001; Trafton & Trickett, 2001). Implementation of cognitively principled design (Bennett, 1999) may been seen in CBTs that incorporate aspects of editorial control, creating a test interface similar to P&Ps (i.e., Goldberg & Pedulla, 2002; Luecht et al., 1998; Mason, et al., 2001). In addition to creating a cognitively comparable test experience across modes of administration, the current study further examined the cognitive factor of loading on working memory (WM). Theories of working memory state that information is temporarily maintained and manipulated and that the capability for such actions is differentially affected by task difficulty and working memory capacity, or ability (Baddeley, 2002). Therefore, as items load differently due to item difficulty or examinee's working memory ability, differences in performance will be evident.

Bennett's (1998) generational theories for computer-based test design and progression can be applied not only to CBTs but to test development and interpretation in general. As working memory is increasingly linked to performance and general intelligence (Kyllonen & Christal, 1990), theories of user-interface design (UID) must be generally applied, both in the design of computer- and paper-based test forms.

FOOTNOTES

[1] Many studies have assumed comparability and statistical equivalence of working memory tests, and thus working memory ability, across P&P and CBT conditions (Kyllonen & Christal, 1990; Miyake, Friedman, Rettinger, Shah, & Hegarty, 2001). These studies simply computerize previously paper-based cognitive tests (i.e., ASVAB subtests).

Recently, working memory ability, specifically regarding the VSSP, has been shown to be stable across administration conditions. Strong and Smith (2002) computerized Vandenberg's Mental Rotation Test (1978) and administered both the computer- and paper-based versions to 30 university students. The Mental Rotation Test is a test of visuo-spatial ability in which students must choose from among four options the three-dimensional diagram that represents a rotation of the target diagram. Randomly assigned to two groups, paper (A) or computer (B) presentation first, neither group evidenced a significant difference in performance, $t_A(14) = 1.372$; $t_B(14) = 1.584$. Both groups, however, showed strong correlations between conditions, $r_A = 0.687$; $r_B = 0.777$.

[2] Syllogistic performance and Paper Folding Test performance were found to be moderately correlated, $r = 0.37$, $p<.05$. Performance on the Paper Folding Test was not found to significantly differ across modes of administration, $t(207) = 1.61$, $p = 0.109$ (Table 10).

REFERENCES

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1986). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

American Psychological Association. (1986). *Guidelines for Computer-based Tests and Interpretation*. Washington, DC: Author.

Baddeley, A. D. (1986). *Working memory*. New York: Oxford University Press.

Baddeley, A. D. (2002). Is working memory still working? *European Psychologist, 7*(2), 85-97.

Baddeley, A. D., & Hitch, G. (1974). Working memory. In G.H. Bower (Ed.), *The psychology of learning and motivation: Advances in research and theory* (pp. 47-89). New York: Academic Press.

Bennett, R. E. (1998). *Reinventing Assessment. Speculations on the Future of Large-Scale Educational Testing. A Policy Information Perspective.* Princeton, NJ: Educational Testing Service.

Bennett, R. E. (1999). Using New Technology To Improve Assessment. *Educational Measurement: Issues & Practice, 18*(3), 5-12.

Booth, J. (1991). The key to valid computer-based testing: The user interface. *European Review of Applied Psychology, 41*(4), 281-293.

Booth, J. F. (1998a). The user interface in computer-based selection and assessment: Applied and theoretical problematics of an evolving technology. *International Journal of Selection & Assessment, 6*(2), 61-81.

Booth, J. F. (1998b). Uses of PC technology in selection and assessment. *International Journal of Selection & Assessment, 6*(2), 57-60.

Bucciarelli, M., & Johnson-Laird, P. N. (1999). Strategies in syllogistic reasoning. *Cognitive Science, 23*(3), 247-303.

Cary, M., & Carlson, R. A. (1999). External support and the development of problem-solving routines. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 25*(4), 1053-1070.

Cary, M., & Carlson, R. A. (2001). Distributing working memory resources during problem solving. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 27*(3), 836-848.

Chalmers, P. A. (2003). The role of cognitive theory in human-computer interface. *Computers in Human Behavior, 19*, 593-607.

Citizenship and Immigration Canada (n.d.). *National Occupation Classification List*. Retrieved July 2, 2003, from http://www.cic.gc.ca/english/skilled/qual-2-1.html

Dimock, P. H., & Cormier, P. (1991). The Effects of Format Differences and Computer Experience on Performance and Anxiety on a Computer-Administered Test. *Measurement & Evaluation in Counselling & Development, 24*(3), 119-126.

Divgi, D. R., & Stoloff, P. H. (1986). *Effect of the medium of administration on ASVAB item response curves.* (Report No. 86-24). Alexandria, VA: Center for Naval Analyses.

Ekstrom, R.B., French, J.W., Harman, H.H., & Dermen, D. (1976). *Kit of factor-referenced cognitive tests*. Princeton, NJ: Educational Testing Service.

Elshout, J. J. (1985, June). *Problem solving and education*. Paper presented at the meeting of the European Association for Research on Learning and Instruction, Leuven, Belgium.

Embretson, S. E. (1999). Cognitive psychology applied to testing. In F.T. Durso, R.S. Nickerson, R.W. Schvaneveldt, S.T. Dumais, D.S. Lindsay, & M.T.H. Chi (Eds.), *Handbook of applied cognition* (pp. 629-660). NY: Wiley & Sons.

Epstein, J., & Klinkenberg, W. D. (2001). From Eliza to Internet: A brief history of computerized assessment. *Computers in Human Behavior, 17*, 295-314.

Fisher, D. L. (1981). A three-factor model of syllogistic reasoning: The study of isolable stages. *Memory & Cognition, 9*(5), 496-514.

Fulcher, G. (1999). Computerizing an English Language Placement Test. *ELT Journal, 53*(4), 289-99.

Gierl, M.J., & Bolt, D.M. (2001). Illustrating the use of nonparametric regression to assess differential item and bundle functioning among multiple groups. *International Journal of Testing, 1*(3&4), 249-270.

Gilhooly, K. J., Logie, R. H., Wetherick, N. E., & Wynn, V. E. (1993). Working memory and strategies in syllogistic-reasoning tasks. *Memory and Cognition, 21*(1), 115-124.

Gilhooly, K. J, Logie, R. H, & Wynn, V. (1999). Syllogistic reasoning tasks, working memory, and skill. *European Journal of Cognitive Psychology, 11*(4), 473-498.

Gilhooly, K. J., Logie, R. H., & Wynn, V. E. (2002). Syllogistic reasoning tasks and working memory: Evidence from sequential presentation of premises. *Current Psychology, 21*(2), 111-120.

Goldberg, A. L., & Pedulla, J. J. (2002). Performance differences according to test mode and computer familiarity on a practice Graduate Record Exam. *Educational & Psychological Measurement, 62*(6), 1053-1067.

Greaud, V. A., & Green, B. F. (1986). Equivalence of conventional and computer presentation of speed tests. *Applied Psychological Measurement, 10*(1), 23-34.

Green, B. F. (1991). Computer-based adaptive testing in 1991. *Psychology & Marketing, 8*(4), 243-257.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of Item Response Theory*. Newbury Park, CA: Sage.

Hegarty, M., & Steinhoff, K. (1997). Individual differences in use of diagrams as external memory in mechanical reasoning. *Learning & Individual Differences, 9*(1), 19-42.

Honaker, L. M. (1988). The equivalency of computerized and conventional MMPI administration: A critical review. *Clinical Psychology Review, 8*(6), 561-577.

Huff, K. L., & Sireci, S. G. (2001). Validity issues in computer-based testing. *Educational Measurement: Issues & Practice, 20*(3), 16-25.

Intons-Peterson, M. J. (1993). External memory aids and their relation to memory. In C. Izawa (Ed.), *Cognitive psychology applied* (pp. 135-158). Hillsdale, NJ: Lawrence Erlbaum Associates.

Intons-Peterson, M. J., & Fournier, J. (1986). External and internal memory aids: When and how do we use them? *Journal of Experimental Psychology: General, 115*, 276-280.

Johnson-Laird, P. N. (1983). *Mental Models*. Cambridge: Cambridge University Press.

Johnson-Laird, P. N., & Bara, B. G. (1984). Syllogistic inference. *Cognition, 16*(1), 1-61.

Johnson-Laird, P. N., & Byrne, R. M. J. (1991). *Deduction*. Hove, UK: Lawrence Erlbaum Associates, Inc.

Kalyuga, S., Ayres, P., Chandler, P., & Sweller, J. (2003). The expertise reversal effect. *Educational Psychologist, 38*(1), 23-31.

Kyllonen, P. C., & Christal, R. E. (1990). Reasoning ability is (little more than) working-memory capacity? *Intelligence, 14*(4), 389-433.

Lee, J. A., Moreno, K. E., & Sympson, J. B. (1986). The effects of mode of test administration on test performance. *Educational & Psychological Measurement, 46*(2), 467-474.

Lord, F. M. (1980). *Applications of Item Response Theory to Practical Testing Problems*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*: Harvard Book List.

Luecht, R. M., Hadadi, A., Swanson, D. B., & Case, S. M. (1998). A comparative study of a comprehensive basic sciences test using paper-and-pencil and computerized formats. *Academic Medicine, 73*(Suppl 10), S51-S53.

Mason, B. J., Patry, M., & Bernstein, D. J. (2001). An examination of the equivalence between non-adaptive computer-based and traditional testing. *Journal of Educational Computing Research, 24*(1), 29-39.

McDonald, A. S. (2002). The Impact of Individual Differences on the Equivalence of Computer-Based and Paper-and-Pencil Educational Assessments. *Computers & Education, 39*(3), 299-312.

Mead, A. D., & Drasgow, F. (1993). Equivalence of computerized and paper-and-pencil cognitive ability tests: A meta-analysis. *Psychological Bulletin, 114*(3), 449-458.

Miyake, A., Friedman, N. P., Rettinger, D. A., Shah, P., & Hegarty, M. (2001). How are visuospatial working memory, executive functioning, and spatial abilities related? A latent-variable analysis. *Journal of Experimental Psychology: General, 130*(4), 621-640.

Nandakumar, R. (1993). Simultaneous DIF amplification and cancellation: Shealy-Stouts test for DIF. *Journal of Educational Measurement, 16*, 159-176.

Neuman, G., & Baydoun, R. (1998). Computerization of Paper-and-Pencil Tests: When Are They Equivalent? *Applied Psychological Measurement, 22*(1), 71-83.

Noyes, J. M., & Garland, K. J. (2003).Solving the Tower of Hanoi: does mode of presentation matter? *Computers in Human Behavior, 19*(5), 579-592.

Pellegrino, J. W. (1995). Technology in support of critical thinking. *Teaching of Psychology, 22*(1), 11-12.

Pellegrino, J. W., Baxter, G. P., & Glaser, R. (1999). Addressing the "two disciplines" problem: Linking theories of cognition and learning with assessment and instructional practice. *Review of Research in Education, 24*, 307-353.

Ramsay, J.O. (1991). Kernel smoothing approaches to nonparametric item characteristic curve estimation. *Psychometrika, 56*, 611-630.

Ramsay, J.O. (2000). *TESTGRAF: A program for the graphical analysis of multiple-choice test and questionnaire data* [Computer program and manual]. Retrieved from http://www.psych.mcgill.ca/faculty/ramsay/ramsay.html.

Reinking, D., & Schreiner, R. (1985). The Effects of Computer-Mediated Text on Measures of Reading Comprehension and Reading Behavior. *Reading Research Quarterly, 20*(5), 536-52.

Russell, M. (1999). Testing on Computers: A Follow-up Study Comparing Performance on Computer and on Paper. *Education Policy Analysis Archives, 7*(20). Retrieved July 2, 2003 from http://epaa.asu.edu/epaa/v7n20/.

Russell, M., & Haney, W. (1997). Testing Writing on Computers: An Experiment Comparing Student Performance on Tests Conducted via Computer and via Paper-and-Pencil. *Education Policy Analysis Archives, 5*(3). Retrieved July 2, 2003 from http://epaa.asu.edu/epaa/v5n3/.

Schönpflung, W. (1986). The trade-off between internal and external information storage. *Journal of Memory and Language, 25*, 657-675.

Segall, D. O., & Moreno, K. E. (1999). Development of the Computerized Adaptive Testing version of the Armed Services Vocational Aptitude Battery. F. Drasgow, & J. B. Olson-Buchanan (Eds.), *Innovations in computerized assessment* (pp. 35-65).

Snow, R. E., & Lohman, D. F. (1989). Implications of cognitive psychology for educational measurement. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 263-331). New York, Macmillan.

Spray, J. A., Ackerman, T. A., Reckase, M. D., & Carlson, J. E. (1989). Effect of the medium of item presentation on examinee performance and item characteristics. *Journal of Educational Measurement, 26*(3), 261-271.

Steinberg, L., Thissen, D., & Wainer, H. (1990). Validity. H. Wainer (Ed.), *Computerized Adaptive Testing: A Primer* (pp. 187-231). Hillsdale, NJ: Erlbaum.

Sternberg, R. J. (1984). What cognitive psychology can (and cannot) do for test development. In B.S. Plake (Ed.), *Social and Technical Issues in Testing* (Vol. 1, pp. 39-60). Hillsdale, NJ: Lawrence Erlbaum.

Strong, S., & Smith, R. (2002). The development of a computerized version of Vandenberg's Mental Rotation Test and the effect of visuo-spatial working memory Loading. *Engineering Design Graphics Journal, 66*(2), 6-16.

Trafton, J. G., & Trickett, S. B. (2001). Note-taking for self-explanation and problem solving. *Human-Computer Interaction, 16*(1), 1-38.

Van de Vijver, F. J. R., & Harsveldt, M. (1994). The incomplete equivalence of the paper-and-pencil and computerized versions of the General Aptitude Test Battery. *Journal of Applied Psychology, 79*(6), 852-859.

Vandenberg, S. G., & Kuse, A. R. (1978). Mental rotations, a group test of three-dimensional spatial visualization. *Perceptual & Motor Skills, 47*(2), 599-604.

Vispoel, W. P., Rocklin, T. R. Wang, T. (1994). Individual differences and test administration procedures: A comparison of fixed-item, computerized-adaptive, self-adapted testing. *Applied Measurement in Education, 7*(1), 53-79.

Wainer, H. (2000). *Computer Adaptive Testing: A Primer* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Wang, T. Y., & Kolen, M. J. (2001). Evaluating Comparability in Computerized Adaptive Testing: Issues, Criteria and an Example. *Journal of Educational Measurement, 38*(1), 19-49.

Wise, S.L., & Plake, B.S. (1989). Research on administering tests via computers. *Educational Measurement: Issues and Practice, 8*, 5-10.

Yeung, A. S. (1999). Cognitive load and learner expertise: split attention and redundancy effects in reading comprehension tasks with vocabulary definitions. *Journal of Experimental Education, 67*(3), 197-221.

TABLES

Table 1
Syllogistic Categories Selected from the National Occupation Classification List

Painters	Welders
Politicians	Entrepreneurs
Bankers	Artists
Athletes	Gardeners*
Nurses	Bookkeepers*
Secretaries	Musicians*
Doctors*	Professors
Dancers*	Accountants
Divers	Shoppers
Teachers*	Surgeons
Golfers*	Researchers
Carpenters*	Pilots
Managers	Architects
Singers	Chiropractors
Hairdressers	Dentists
Gardeners*	Engineers
Bookkeepers*	Journalists
Musicians*	Pharmacists
Teachers*	Bricklayers
Golfers*	Cooks
Carpenters*	Firefighters
Swimmers	Meteorologists
Janitors	Optometrists
Plumbers	Editors
Librarians	Biologists
Travellers	Barbers
Counsellors	Astronauts
Doctors*	Clerks
Dancers*	Cabbies
Drivers	Couriers
Waiters	Authors
Farmers	Inspectors
Electricians	Pipefitters
Photographers	Tailors

*: Some of the occupations were used more than once.

Table 2:
Demographic Proportions of Experimental Sample

		Full (n = 209)	Paper (n = 110)		Computer (n = 99)	
			No Aid (n = 58)	Aid (n = 52)	No Aid (n = 50)	Aid (n = 49)
Faculty	Arts	39.71	39.66	38.46	36.00	44.90
	Sciences	48.33	51.72	44.23	54.00	42.86
	Other	11.96	8.62	17.31	10.00	12.24
Year of Study	First	55.50	60.34	50.00	58.00	53.06
	Second	30.14	27.59	34.62	32.00	26.53
	Third	9.57	5.17	11.54	4.00	18.37
	Fourth	2.39	5.17	1.92	0.00	2.04
	Other	2.39	1.72	1.92	6.00	0.00
Age	Under 18	0.48	0.00	1.92	0.00	0.00
	18-20	79.90	82.76	73.08	86.00	77.55
	21-24	14.83	13.79	15.38	10.00	20.41
	25-30	3.35	3.45	7.69	2.00	0.00
Gender	Female	67.46	60.34	75.00	60.00	75.51
	Male	32.54	39.66	25.00	40.00	24.49

Note: values are proportions of the total

Table 3:

Demographic Proportions of Five-Point Computer Likert Self-Assessment

				Paper (n = 110)		Computer (n = 99)	
			Full (n = 209)	No Aid (n = 58)	Aid (n = 52)	No Aid (n = 50)	Aid (n = 49)
Comfort		mean (s.d.)	4.02 (0.71)	4.02 (0.74)	3.92 (0.71)	4.10 (0.68)	4.04 (0.71)
with	Not at all		0.00	0.00	0.00	0.00	0.00
Computers	Slightly		2.39	3.45	3.85	0.00	2.04
	Somewhat	proportion	16.75	15.52	17.31	18.00	16.33
	Comfortable		57.42	56.90	61.54	54.00	57.14
	Very		23.44	24.14	17.31	28.00	24.49
Word		mean (s.d.)	4.17 (0.73)	4.24 (0.84)	4.00 (0.71)	4.18 (0.72)	4.24 (0.60)
Processing	None		0.48	1.72	0.00	0.00	0.00
Skill	Poor		0.96	1.72	1.92	0.00	0.00
	Fair	proportion	13.88	10.34	19.23	18.00	8.16
	Good		50.72	43.10	55.77	46.00	59.18
	Excellent		33.97	43.10	23.08	36.00	32.65
Email		mean (s.d.)	4.35 (0.69)	4.33 (0.85)	4.27 (0.63)	4.38 (0.67)	4.42 (0.58)
Skill	None		0.48	1.72	0.00	0.00	4.08
	Poor		0.96	3.45	0.00	0.00	48.98
	Fair	proportion	6.70	3.45	9.62	10.00	44.90
	Good		46.89	43.10	53.85	42.00	97.96
	Excellent		44.50	48.28	36.54	48.00	2.04

Table 3 (cont.):

Demographic Proportions of Five-Point Computer Likert Self-Assessment

				Paper (n = 110)		Computer (n = 99)	
			Full (n = 209)	No Aid (n = 58)	Aid (n = 52)	No Aid (n = 50)	Aid (n = 49)
Games		mean (s.d.)	3.40 (1.09)	3.53 (1.19)	3.17 (1.04)	3.50 (1.02)	3.39 (1.10)
Skill	None		4.31	6.90	3.85	4.00	2.04
	Poor		16.75	13.79	21.15	10.00	22.45
	Fair	proportion	31.10	20.69	42.31	34.00	28.57
	Good		30.14	36.21	19.23	36.00	28.57
	Excellent		17.70	22.41	13.46	16.00	18.37
Searching		mean (s.d.)	3.91 (0.74)	3.90 (0.85)	3.85 (0.70)	3.98 (0.68)	3.92 (0.73)
the	None		0.48	1.72	0.00	0.00	0.00
Web	Poor		2.87	3.45	1.92	2.00	4.08
Skill	Fair	proportion	21.05	20.69	26.92	18.00	18.37
	Good		56.46	51.72	55.77	60.00	59.18
	Excellent		19.14	22.41	15.38	20.00	18.37

Table 4
Item p-Values Across Experimental Conditions

	Model Type			Computer	Paper	Aid	No-Aid	Full
Item	One	Multiple	NVC	(n = 99)	(n = 110)	(n = 101)	(n = 108)	(n = 209)
1	x			0.88	0.93	0.90	0.90	0.90
2	x			0.77	0.81	0.82	0.76	0.79
3		x		0.12	0.15	0.12	0.15	0.14
4		x		0.43	0.49	0.46	0.46	0.46
5			x	0.54	0.61	0.60	0.54	0.57
6			x	0.66	0.64	0.62	0.68	0.65
7	x			0.72	0.85	0.80	0.77	0.79
8	x			0.85	0.85	0.83	0.87	0.85
9		x		0.37	0.47	0.36	0.48	0.43
10		x		0.27	0.36	0.35	0.29	0.32
11			x	0.52	0.62	0.57	0.56	0.57
12			x	0.52	0.64	0.58	0.57	0.58
13	x			0.87	0.82	0.90	0.79	0.84
14	x			0.82	0.88	0.89	0.81	0.85
15		x		0.38	0.36	0.37	0.38	0.37
16		x		0.32	0.39	0.37	0.34	0.36
17			x	0.67	0.76	0.71	0.72	0.72
18			x	0.45	0.51	0.48	0.48	0.48
19	x			0.73	0.70	0.68	0.74	0.71
20	x			0.72	0.72	0.73	0.70	0.72
21		x		0.28	0.31	0.29	0.30	0.30
22		x		0.29	0.33	0.34	0.29	0.31
23			x	0.66	0.77	0.74	0.69	0.72
24			x	0.68	0.80	0.72	0.75	0.74
Mean	0.81	0.34	0.63	0.56	0.62	0.59	0.58	0.59
S.D.	0.07	0.10	0.09	0.22	0.22	0.22	0.21	0.22

Table 5
Intercorrelations Between Item p-values Across Conditions

			Correlations (n = 24)			
Condition	Mean	Std. Dev.	Computer	Paper	Aid	No-Aid
Computer	0.56	0.22		0.97	0.99	0.99
Paper	0.62	0.22			0.99	0.99
Aid	0.59	0.22				0.97
No-Aid	0.58	0.21				

Note: all correlations are significant at $p<.05$

Table 6
Syllogistic Performance

Mode of Administration	Memory Aid Condition	Mean	Standard Deviation	n
Paper	No Aid	14.72	4.36	58
	Aid	14.87	4.68	52
	Total	14.79	4.49	110
Computer	No Aid	13.34	4.35	50
	Aid	13.67	3.93	49
	Total	13.51	4.13	99
Total	No Aid	14.08	4.39	108
	Aid	14.29	4.35	101
	Total	14.18	4.36	209

Table 7

Analysis of Variance for Total Syllogistic Performance (n = 209)

Source	df	F	p
Mode of Administration (M)	1	4.58*	0.03
Memory Aid (A)	1	0.16	0.69
M x A	1	0.03	0.87
Error	205	(18.85)	

*: $p<.05$

Table 8
Syllogistic Performance Across Modes and Memory Aid Use

Mode of Administration	Memory Aid Use	Mean	Standard Deviation	n
Paper	No Use	13.83	4.53	35
	Use	17.00	4.36	17
	Total	14.87	4.68	52
Computer	No Use	12.39	2.91	28
	Use	15.38	4.50	21
	Total	13.67	3.93	49
Total	No Use	13.19	3.93	63
	Use	16.11	4.45	38
	Total	14.29	4.35	101

Note: maximum score = 24

Table 9

Analysis of Variance for Syllogistic Performance Across Modes and Memory Aid Use (n = 101)

Source	df	F	p
Mode of Administration (M)	1	3.24**	0.07
Memory Aid Use (U)	1	13.19*	0.00
M x U	1	0.01	0.91
Error	97	(16.85)	

*: $p<.05$

**: $p<.10$

Table 10
Paper Folding Test Performance Across Modes of Administration

Mode of Administration	Mean	Standard Deviation	n
Paper	15.18	4.09	110
Computer	14.28	3.98	99
Total	14.75	4.05	209

Note: maximum score = 20

Table 11

Syllogistic Performance Across Modes and Memory Aid Given Visuo-Spatial Ability

Visuo-Spatial Ability	Mode of Administration	Memory Aid Condition	Mean	Standard Deviation	n
LOW	Paper	No Aid	14.29	4.52	31
		Aid	13.38	4.84	26
		Total	13.88	4.65	57
	Computer	No Aid	12.79	3.48	33
		Aid	13.00	3.85	31
		Total	12.89	3.63	64
	Total	No Aid	13.52	4.06	64
		Aid	13.18	4.29	57
		Total	13.36	4.16	121
HIGH	Paper	No Aid	15.22	4.19	27
		Aid	16.35	4.08	26
		Total	15.77	4.14	53
	Computer	No Aid	14.41	5.64	17
		Aid	14.83	3.90	18
		Total	14.63	4.75	35
	Total	No Aid	14.91	4.75	44
		Aid	15.73	4.03	44
		Total	15.32	4.40	88

Note: maximum score = 20

Table 12

Analysis of Variance for Syllogistic Performance Across Modes and Memory Aid Given Visuo-Spatial Ability (n = 209)

Visuo-Spatial Ability	Source	df	F	p
LOW	Mode of Administration (M)	1	1.54	0.22
	Memory Aid (A)	1	0.21	0.65
	M x A	1	0.54	0.46
	Error	117	(17.35)	
HIGH	Mode of Administration (M)	1	1.46	0.23
	Memory Aid (A)	1	0.64	0.42
	M x A	1	0.13	0.72
	Error	84	(19.51)	

Table 13
Syllogistic Performance Across Modes and Memory Aid Use Given Visuo-Spatial Ability (n = 101)

Visuo-Spatial Ability	Mode of Administration	Memory Aid Use	Mean	Standard Deviation	n
LOW	Paper	No Use	11.59	3.92	17
		Use	16.78	4.76	9
		Total	13.38	4.84	26
	Computer	No Use	12.00	2.92	21
		Use	15.10	4.82	10
		Total	13.00	3.85	31
	Total	No Use	11.82	3.36	38
		Use	15.89	4.74	19
		Total	13.18	4.29	57
HIGH	Paper	No Use	15.94	4.09	18
		Use	17.25	4.17	8
		Total	16.35	4.08	26
	Computer	No Use	13.57	2.76	7
		Use	15.64	4.41	11
		Total	14.83	3.90	18
	Total	No Use	15.28	3.87	25
		Use	16.32	4.27	19
		Total	15.73	4.03	44

Note: maximum score = 24

Table 14

Analysis of Variance for Syllogistic Performance Across Modes and Memory Aid Use Given Visuo-Spatial Ability (n = 101)

Visuo-Spatial Ability	Source	df	F	p
LOW	Mode of Administration (M)	1	0.33	0.57
	Memory Aid Use (U)	1	14.22*	0.00
	M x-U	1	0.90	0.35
	Error	53	(15.22)	
HIGH	Mode of Administration (M)	1	2.37	0.13
	Memory Aid Use (U)	1	1.70	0.20
	M x U	1	0.09	0.77
	Error	40	(16.17)	

*: $p<.05$

Table 15

Beta Values Across Experimental Conditions (n = 209)

	Reference	P		A			PA					
	Focal	Com		N			ComA		ComN		PN	
Model		β_F	DIF	β_F	DIF	β_R	β_F	DIF	β_F	DIF	β_F	DIF
One		-0.026		0.011		0.062	0.054	B	0.026		-0.012	
		0.068	B	0.000		0.090	-0.021		0.014		0.052	B
		-0.004		0.066	B	0.085	0.105	C	-0.004		0.050	B
		-0.049		-0.013		0.092	0.011		-0.024		-0.036	
		0.128	C	-0.061	B	0.109	-0.059	B	0.123	C	0.030	
		0.084	B	0.030		0.112	-0.043		0.047		0.068	B
		-0.051	B	-0.045		0.110	-0.048		-0.022		-0.155	C
		0.007		-0.038		0.092	0.039		0.105	C	-0.031	
Multiple		-0.034		0.003		0.052	0.030		0.013		-0.007	
		-0.004		0.004		0.076	-0.022		-0.029		0.017	
		-0.099	B	0.034		0.145	0.013		-0.190	C	-0.014	
		0.028		0.033		0.092	0.071	B	0.042		0.065	B
		0.004		-0.057	B	0.143	-0.097	B	0.037		-0.124	C
		0.041		-0.033		0.152	-0.185	C	-0.120	C	0.026	
		0.034		0.001		0.106	-0.068	B	-0.083	B	0.045	B
		0.064	B	-0.042		0.143	-0.016		0.070	B	0.094	B
NVC		0.045		0.006		0.102	0.045		0.117	C	0.034	
		-0.064	B	-0.090	B	0.119	-0.093	B	-0.092	B	-0.136	C
		-0.050	B	0.005		0.081	-0.032		-0.040		-0.045	
		-0.003		0.036		0.065	0.013		-0.035		0.010	
		-0.054	B	0.009		0.078	0.055	B	-0.010		-0.078	B
		-0.027		0.057	B	0.108	0.093	B	0.039		0.055	B
		0.024		0.034		0.100	0.059	B	0.047		0.099	B
		-0.064	B	0.047		0.077	0.093	B	-0.031		-0.010	

Note: P=Paper, Com=Computer, A=Aid, N=No-Aid

Table 16

Beta Values Across Mode of Administration and Use of Memory Aid (n = 101)

	Reference	P		U			PU					
	Focal	Com		X			ComU		ComX		PX	
Model		β_F	DIF	β_F	DIF	β_R	β_F	DIF	β_F	DIF	β_F	DIF
One		0.040		-0.013		0.041	-0.009		0.010		-0.049	
		0.020		-0.033		0.119	-0.036		0.020		-0.053	B
		0.076	B	0.034		0.084	-0.021		0.040		-0.044	
		-0.009		0.087	B	0.111	0.012		0.059	B	0.078	B
		0.119	C	0.077	B	0.199	0.005		0.192	C	-0.031	
		0.080	B	0.189	C	0.263	0.022		0.300	C	0.170	C
		-0.014		0.036		0.087	0.000		0.006		0.024	
		0.021		0.044		0.123	0.064	B	0.112	C	0.010	
Multiple		0.002		0.030		0.132	0.143	C	0.036		0.080	B
		0.010		0.157	C	0.320	0.235	C	0.237	C	0.333	C
		-0.180	C	0.071	B	0.170	0.017		-0.126	C	0.104	C
		-0.003		0.128	C	0.323	0.341	C	0.196	C	0.280	C
		0.055	B	-0.030		0.153	0.046		0.101	C	0.045	
		-0.133	C	0.038		0.159	0.076	B	-0.062		0.157	C
		-0.062		-0.133	C	0.224	-0.173	C	-0.187	C	-0.980	B
		0.023		-0.059	B	0.195	0.164	C	-0.121	C	-0.107	C
NVC		0.159	C	-0.025		0.183	-0.152	C	0.047		-0.196	C
		-0.127	C	-0.121	C	0.243	-0.198	C	-0.215	C	-0.175	C
		-0.072	B	-0.013		0.171	-0.154	C	-0.146	C	-0.054	B
		-0.071	B	-0.021		0.117	-0.104	C	0.012		0.029	
		0.037		-0.038		0.112	-0.029		-0.075	B	-0.088	B
		0.048		-0.141	C	0.176	-0.101	C	-0.087	B	-0.111	C
		0.047		-0.224	C	0.270	-0.061	B	-0.251	C	-0.271	C
		-0.068	B	-0.041		0.121	-0.087	B	-0.101	C	-0.032	

Note: P=Paper, Com=Computer, U=Use, X=Non-Use

Table 17

Beta Values Across Experimental Conditions According to Low Visuo-Spatial Ability (n = 121)

	Reference	P		A			PA					
	Focal	Com		N			ComA		ComN		PN	
Model		β_F	DIF	β_F	DIF	β_R	β_F	DIF	β_F	DIF	β_F	DIF
One		-0.015		0.010		0.052	-0.018		0.021		0.014	
		0.086	B	-0.068	B	0.114	-0.026		0.000		-0.141	C
		0.010		0.043		0.061	0.026		0.038		0.051	B
		-0.092	B	-0.079	B	0.131	-0.077	B	-0.175	C	-0.013	
		0.054	B	-0.102	C	0.103	0.060	B	-0.109	C	-0.108	C
		0.097	B	0.029		0.115	0.061	B	0.132	C	-0.033	
		-0.018		-0.079	B	0.110	-0.030		-0.021		-0.083	B
		0.074	B	-0.127	C	0.110	0.163	C	-0.067	B	0.019	
Multiple		-0.032		0.000		0.064	0.027		-0.017		0.061	B
		0.070	B	-0.086	B	0.104	0.115	C	-0.078	B	0.013	
		-0.173	C	0.045		0.140	-0.092	B	-0.158	C	0.108	C
		-0.041		0.093	B	0.139	0.117	C	-0.034		0.184	
		-0.011		-0.108	C	0.126	0.132	C	-0.132	C	-0.008	
		0.007		-0.059	B	0.127	-0.116	C	-0.028		-0.120	C
		-0.370		0.084	B	0.146	-0.045		-0.130	C	-0.119	C
		0.031		-0.064	B	0.125	0.023		-0.047		-0.029	
NVC		0.077	B	0.132	C	0.171	0.053	B	0.243	C	0.082	B
		0.020		-0.025		0.068	0.006		0.006		-0.048	
		-0.057	B	-0.030		0.131	-0.128	C	-0.073	B	-0.078	B
		0.074	B	0.123	C	0.138	-0.020		0.249	C	0.016	
		0.012		0.130	C	0.112	0.027		0.131	C	0.149	C
		-0.053	B	0.200	C	0.131	-0.060	B	0.158	C	0.072	B
		-0.006		0.056	B	0.108	-0.077	B	0.150	C	0.000	
		-0.076	B	0.050	B	0.098	-0.121	C	-0.057	B	0.011	

Note: P=Paper, C=Computer, A=Aid, N=No-Aid

Table 18

Beta Values Across Mode of Administration and Use of Memory Aid According to Low Visuo-Spatial Ability (n = 57)

	Reference	P		U			PU					
	Focal	Com		X			ComU		ComX		PX	
Model		β_F	DIF	β_F	DIF	β_R	β_F	DIF	β_F	DIF	β_F	DIF
One		-0.007		0.027		0.055	-0.007		0.034		0.013	
		-0.026		-0.021		0.169	0.016		-0.012		-0.024	
		0.022		0.081	B	0.143	0.001		0.118	C	0.080	B
		-0.070	B	0.210	C	0.132	-0.075	B	0.071	B	0.120	C
		0.059	B	0.112	C	0.134	-0.024		0.141	C	0.026	
		0.054	B	0.184	C	0.[illegible]	0.001		0.275	C	0.165	C
		-0.030		0.026		0.137	-0.039		-0.107	C	-0.029	
		0.156	C	-0.001		0.148	0.022		0.091	B	-0.038	
Multiple		0.020		-0.005		0.068	0.043		0.021		-0.001	
		0.111	C	0.104	C	0.372	0.217	C	0.318	C	0.414	C
		-0.097	B	0.050	B	0.117	-0.011		0.001		0.147	C
		0.109	C	0.127	C	0.238	0.088	B	0.287	C	0.249	C
		0.136	C	-0.094	B	0.129	0.044		0.119	C	0.010	
		-0.132	C	0.077	B	0.173	0.007		-0.059	B	0.169	C
		-0.047		-0.175	C	0.156	0.014		-0.079	B	-0.114	C
		0.003		-0.108	C	0.170	0.024		-0.149	C	-0.041	C
NVC		0.067	B	0.002		0.175	-0.059	B	0.007		-0.195	C
		0.019		-0.095	B	0.197	-0.067	B	-0.126	C	-0.208	C
		-0.122	C	-0.100	C	0.224	-0.068	B	-0.260	C	-0.190	C
		-0.018		-0.088	B	0.106	0.024		-0.153	C	-0.067	B
		0.026		-0.039		0.109	-0.052	B	-0.051	B	-0.144	C
		-0.050	B	-0.007		0.121	0.026		0.026		0.041	
		-0.067	B	-0.193	C	0.286	-0.103	C	-0.316	C	-0.330	C
		-0.116	C	-0.075	B	0.132	-0.022		-0.198	C	-0.050	B

Note: P=Paper, Com=Computer, U=Use, X=Non-Use

Table 19

Beta Values Across Experimental Conditions According to High Visuo-Spatial Ability (n = 88)

	Reference	P		A			PA					
	Focal	Com		N			ComA		ComN		PN	
Model		β_F	DIF	β_F	DIF	β_R	β_F	DIF	β_F	DIF	β_F	DIF
One		-0.020		0.003		0.113	0.104	C	-0.014		0.112	C
		0.014		0.120	C	0.184	0.085	B	0.141	C	0.200	C
		0.076	B	0.077	B	0.159	0.102	C	0.171	C	0.108	C
		0.015		-0.009		0.084	0.075	B	0.003		0.020	
		0.237	C	-0.008		0.243	0.268	C	0.204	C	0.017	
		0.112	C	-0.046		0.117	0.059	B	0.046		-0.069	B
		-0.210	C	-0.036		0.172	-0.085	B	-0.205	C	0.095	B
		-0.118	C	0.035		0.112	-0.088	B	-0.077	B	0.038	
Multiple		0.003		0.007		0.133	-0.026		0.014		-0.045	
		-0.086	B	0.173	C	0.198	-0.231	C	0.077	B	0.053	B
		0.000		0.068	B	0.234	-0.300	C	0.047		-0.090	B
		0.117	C	0.060	B	0.230	-0.123	C	0.211	C	-0.102	C
		-0.010		-0.040		0.164	-0.099	B	-0.056	B	-0.072	B
		0.134	C	0.051	B	0.171	-0.080	B	0.218	C	-0.061	B
		0.128	C	-0.071	B	0.137	-0.016		0.050	B	-0.185	C
		0.116	C	0.024		0.163	0.118	C	0.154	C	0.057	B
NVC		-0.044		-0.096	B	0.220	0.265	C	-0.123	C	0.028	
		-0.253	C	-0.134	C	0.373	-0.256	C	-0.397	C	-0.221	C
		-0.001		0.004		0.116	0.064	B	-0.035		-0.025	
		-0.124	C	-0.045		0.179	-0.113	C	-0.214	C	-0.067	B
		-0.153	C	-0.086	B	0.189	-0.087	B	-0.228	C	-0.047	
		-0.033		-0.074	B	0.122	0.081	B	-0.070	B	0.027	
		0.092	B	-0.024		0.182	0.219	C	0.036		0.116	C
		0.009		0.047		0.119	0.063	B	0.043		0.111	C

Note: P=Paper, Com=Computer, A=Aid, N=No-Aid

Table 20

Beta Values Across Mode of Administration and Use of Memory Aid According to High Visuo-Spatial Ability (n = 44)

	Reference	P		U			PU					
	Focal	Com		X			ComU		ComX		PX	
Model		β_F	DIF	β_F	DIF	β_R	β_F	DIF	β_F	DIF	β_F	DIF
One		0.104	C	-0.063	B	0.147	-0.027		0.000		-0.156	C
		0.102	C	0.043		0.172	0.024		0.084	B	-0.044	
		0.100	C	-0.018		0.150	0.085	B	0.127	C	0.024	
		0.268	C	-0.022		0.408	0.225	C	0.377	C	0.038	
		0.059	B	0.158	C	0.421	0.105	C	0.379	C	0.226	C
		-0.080	B	0.224	C	0.276	0.034		0.197	C	0.225	C
		0.082	B	-0.060	B	0.266	-0.141	C	-0.052	B	-0.279	C
		-0.091	B	0.125	C	0.176	0.039		0.047		0.193	C
Multiple		-0.030		0.077	B	0.194	0.096	B	0.059	B	0.168	C
		-0.231	C	0.177	C	0.326	-0.213	C	-0.181	C	0.093	B
		-0.303	C	0.080	B	0.568	-0.305	C	-0.501	C	-0.053	B
		-0.148	C	0.055	B	0.402	0.319	C	-0.005		0.395	C
		-0.102	C	0.071	B	0.223	-0.058	B	-0.066	B	0.143	C
		-0.080	B	-0.128	C	0.463	-0.268	C	-0.378	C	-0.241	C
		-0.024		-0.130	C	0.241	0.166	C	-0.178	C	0.121	C
		0.110	C	-0.235	C	0.266	0.200	C	-0.178	C	-0.125	C
NVC		0.268	C	-0.137	C	0.366	-0.040		0.234	C	-0.322	C
		-0.255	C	0.075	B	0.449	-0.191	C	-0.363	C	0.017	
		0.067	B	0.000		0.296	0.095	B	0.129	C	0.052	B
		-0.104	C	0.279	C	0.334	-0.077	B	0.253	C	0.176	C
		-0.087	B	-0.103	C	0.162	0.088	B	-0.103	C	0.055	B
		0.086	B	-0.267	C	0.579	-0.267	C	-0.243	C	-0.569	C
		0.222	C	-0.268	C	0.249	0.171	C	0.121	C	-0.139	C
		0.067	B	0.065	B	0.255	-0.063	B	0.236	C	0.000	

Note: P=Paper, Com=Computer, U=Use, X=Non-Use

Table 21
Direction of DIF Evidenced Across Experimental Conditions, Ability, and Syllogism Type

Ability	Reference Group	Focal Group	One Model	Multiple Model	NVC
Full	P	C	+*		
	A	N			
	PA	CA	+	-**	
		CN	+	-	+
		PN	-	-	-
	A|P	A|C	+	-	+/-
	A|U	A|X	+	+/-	-
	A|PU	A|CU		+/-	-
		A|CX	+	+/-	-
		A|PX	+	+/-	-
Low	P	C		-	
	A	N	-	-	+
	PA	CA	+	+/-	-
		CN	+/-	-	+
		PN	-	+/-	+
	A|P	A|C	+	+/-	-
	A|U	A|X	+	+/-	-
	A|PU	A|CU		+	-
		A|CX	+/-	+/-	-
		A|PX	+	+/-	-
High	P	C	+/-	+	-
	A	N	+	+	-
	PA	CA	+	+/-	+/-
		CN	+/-	+	-
		PN	+	-	+/-
	A|P	A|C	+	+/-	+/-
	A|U	A|X	+	+/-	+/-
	A|PU	A|CU	+/-	+/-	+/-
		A|CX	+	-	+/-
		A|PX	+/-	+/-	+/-

Note: P=Paper, C=Computer, A=Aid, N=No-Aid, U=Use, X=Non-Use

*: + indicates DIF favouring the focal group

**: - indicates DIF favouring the reference group

FIGURES

Figure Captions

Figure 1: Syllogistic performance within the memory aid-provided condition according to ability, mode of administration, and use or non-use of the memory. Dashed lines represent non-use; solid lines, use. Low ability is noted by square end-points, high ability by triangle end-points.

Figure 2: Item Characteristic Curve (ICC) generated by Testgraf. The curvilinear line represents the probability (y-axis) that an examinee of a given total score (x-axis) will answer the item correctly. The solid vertical lines indicate the 95% confidence interval for the ICC while the dashed vertical lines indicate frequencies of the total score: 5%, 25%, 50%, 75%, and 95%.

Figure 3: Sample DIF graph. This graph represents DIF across experimental conditions present in multiple-model syllogisms across overall examinee ability. Negative beta-values indicate performance favouring the reference group (paper, aid, or paper*aid); positive values favour the focal group listed. Within each graph, a dashed line represents C-level DIF, $|\beta_F| \geq 0.10$.

Figures 4 - 21: DIF graphs. These graphs display beta-values evidenced by the eight items of the syllogistic type indicated according to experimental conditions. The reference groups, in order, are paper, aid, and paper*aid for the final three comparisons. Negative beta-values indicate performance favouring the reference group; positive values

favour the focal group. Within each graph, a dashed line represents C-level DIF, $|\beta_F| \geq 0.10$.

Figure B1: Paper Folding Test example. This is a sample of the items appearing on the Paper Folding Test (Ekstrom et al., 1976). Examinees were to imagine a piece of paper being folded then punched with a hole, as shown by the images to the left of the vertical line. From the five images to the right of the vertical line, examinees to select the image that correctly represented the initial image as it would appear if unfolded.

Figure C1: Sample instrument page one. Provides examinees with an instrument overview and instructions on completing the included demographic items.

Figure C2: Sample instrument page two. Provides examinees with instructions regarding the syllogistic reasoning task as well as two example syllogisms.

Figure C3: Sample instrument page three. The first twelve of twenty-four syllogisms are presented.

Figure C4: Sample instrument page four. The second set of twelve syllogisms is presented.

Figure C5: Sample instrument page eight. Examinees are thanked for their participation and given further instructions.

Figure D1: Item Characteristic Curves (ICCs) for the twenty-four syllogisms. These graphs display the probability (y-axis) that an examinee of a given total score (x-axis) will answer the item correctly, shown by the curvilinear line. The solid vertical lines indicate the 95% confidence interval for the ICC while the dashed vertical lines indicate frequencies of the total score: 5%, 25%, 50%, 75%, and 95%.

Figure 1: Syllogistic Performance According to Ability and Memory Aid Usage

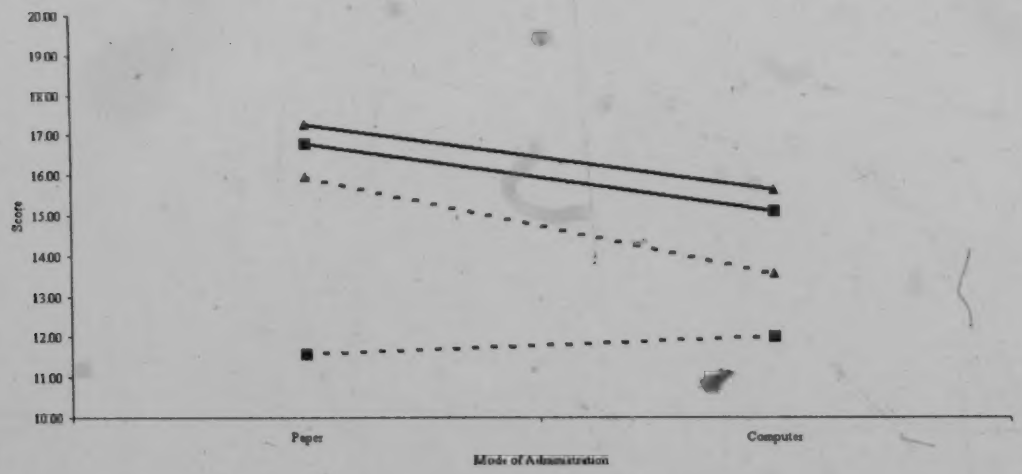


Figure 2: Item Characteristic Curve (ICC) generated by Testgraf

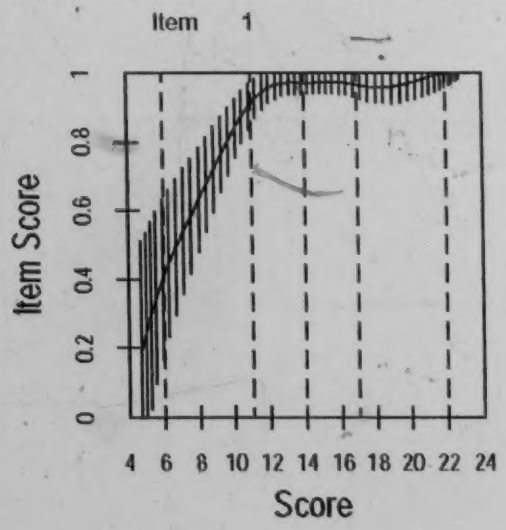


Figure 3: Sample DIF Graph for Overall Ability and Multiple-Model Syllogisms

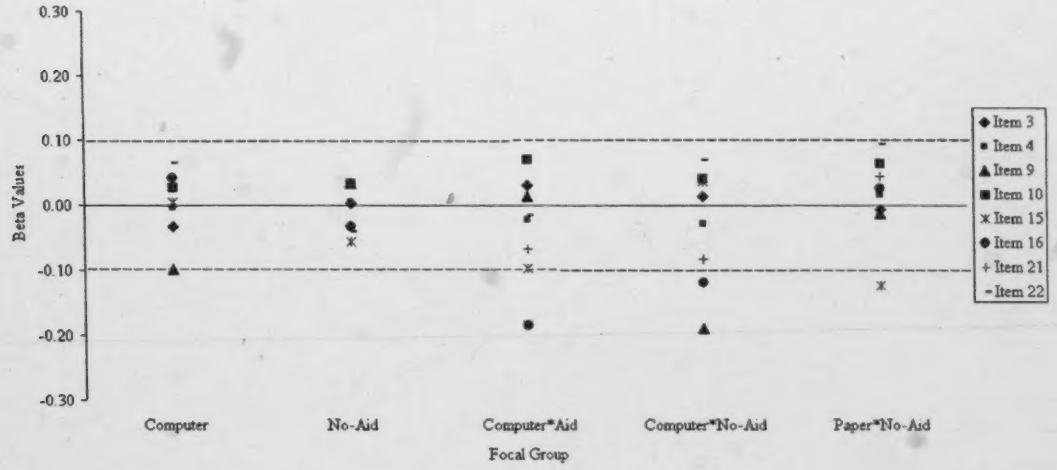


Figure 4: Overall Ability and One-Model Syllogisms

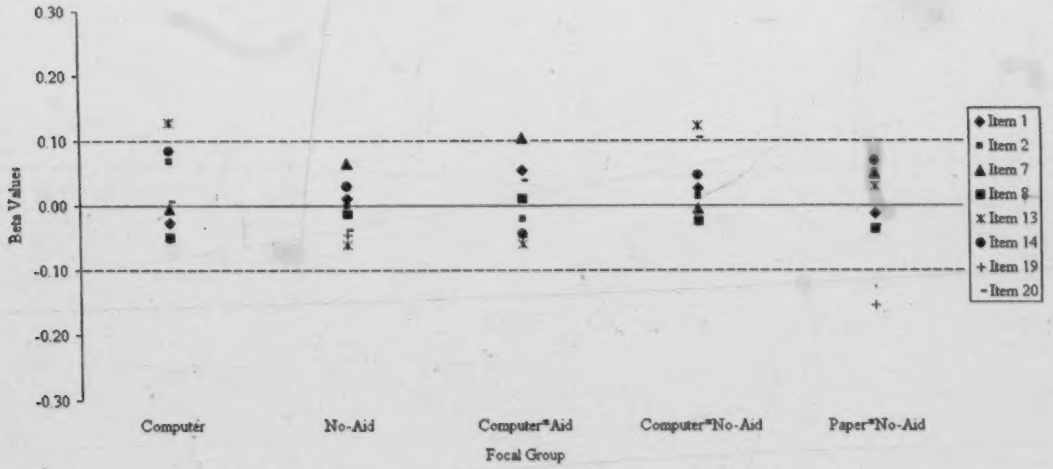


Figure 5: Overall Ability and Multiple-Model Syllogisms

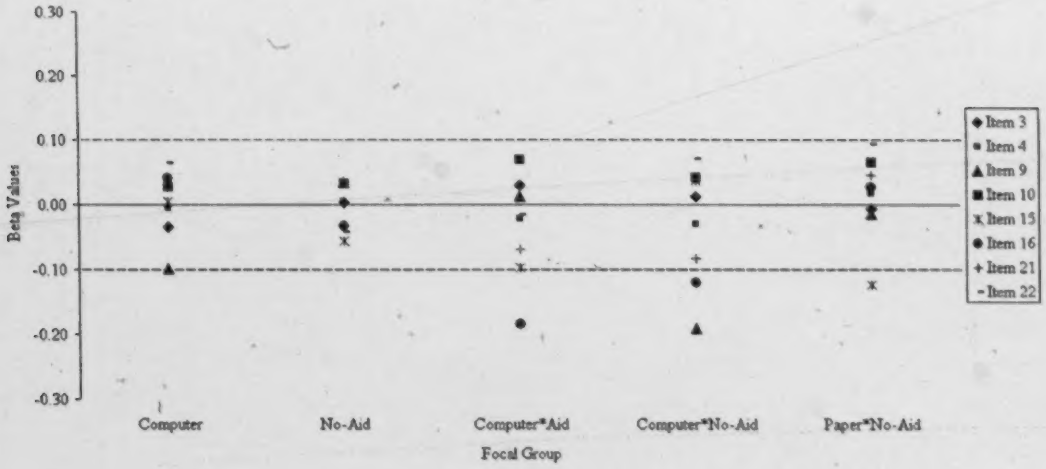


Figure 6: Overall Ability and No Valid Conclusion Syllogisms

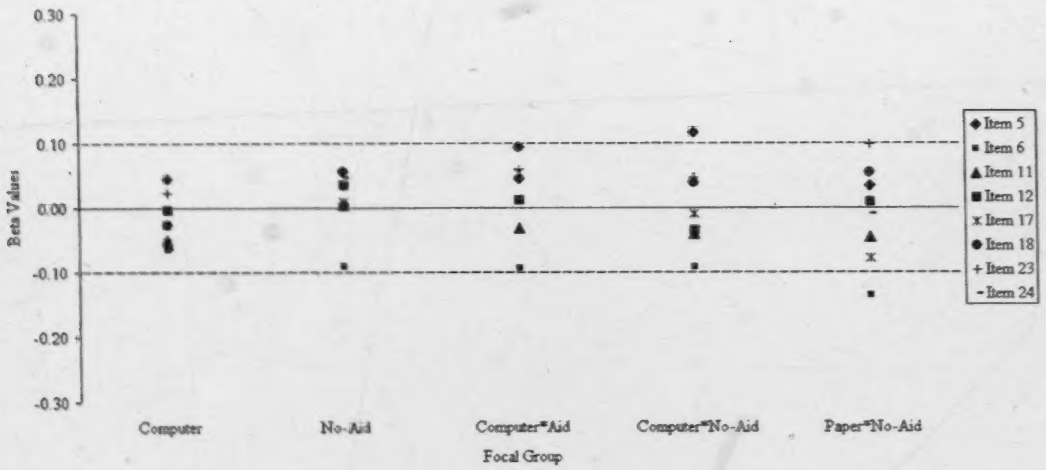


Figure 7: Low Ability and One-Model Syllogisms

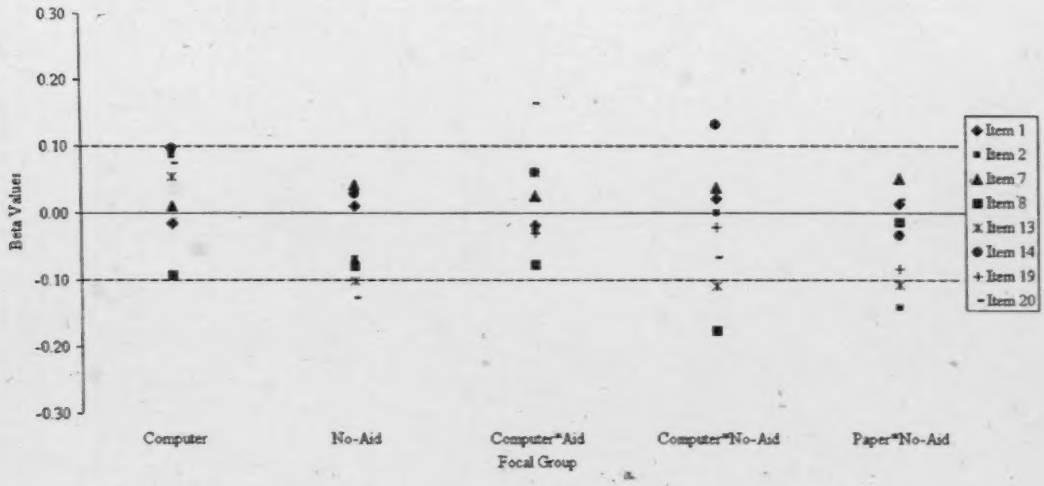


Figure 8: Low Ability and Multiple-Model Syllogisms

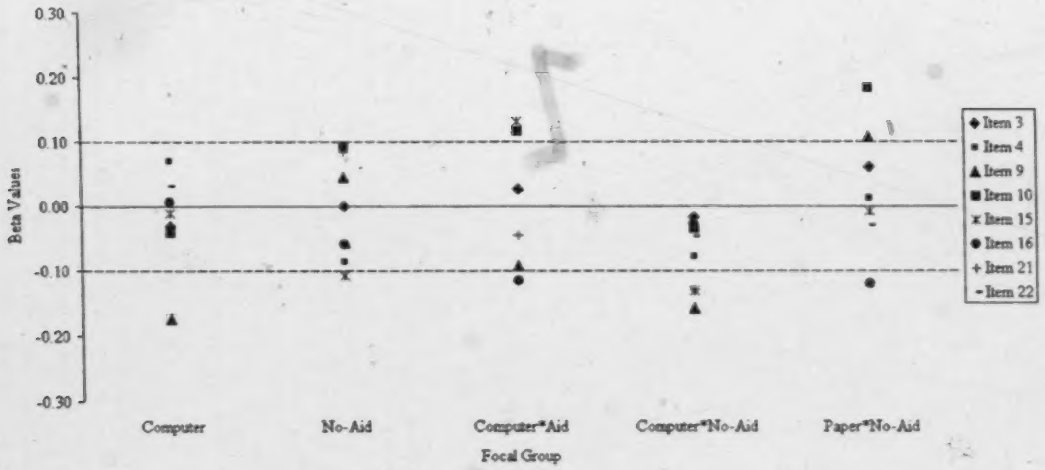


Figure 9: Low Ability and No Valid Conclusion Syllogisms

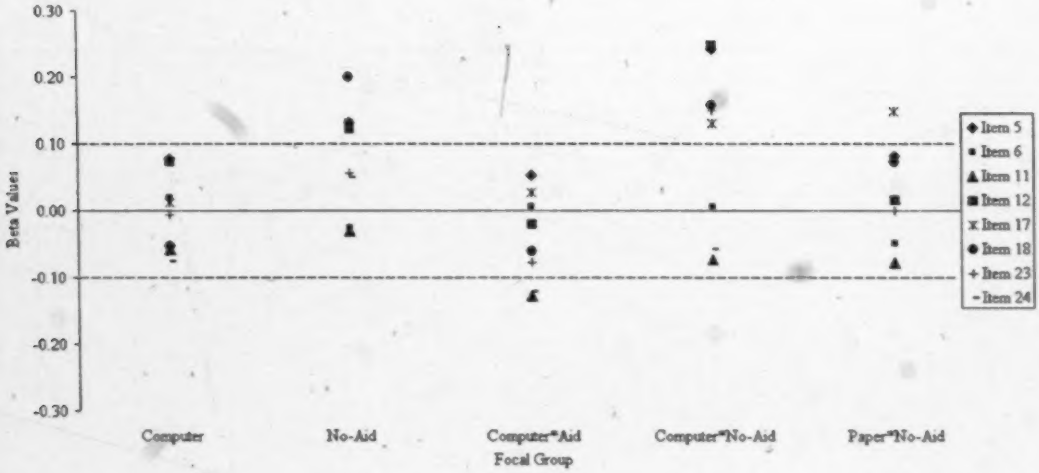


Figure 10: High Ability and One-Model Syllogisms

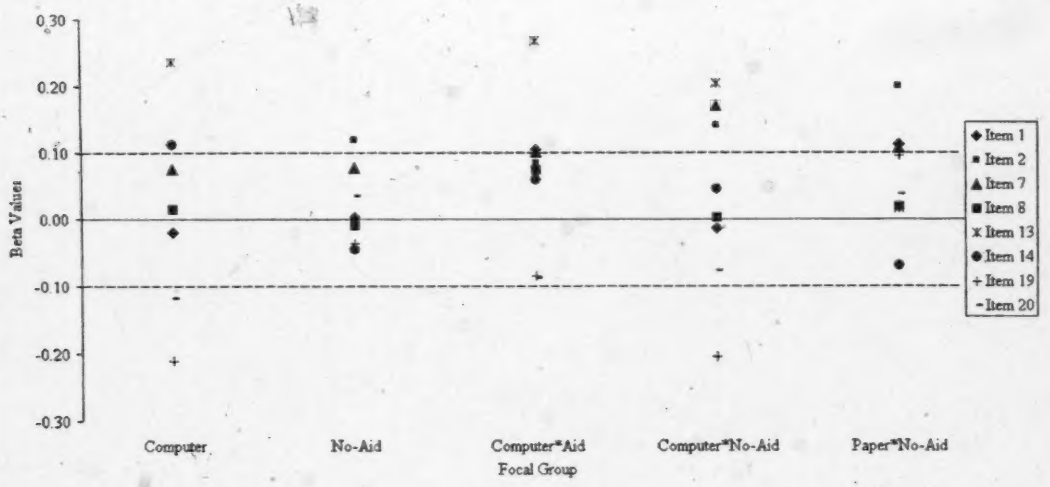


Figure 11: High Ability and Multiple-Model Syllogisms

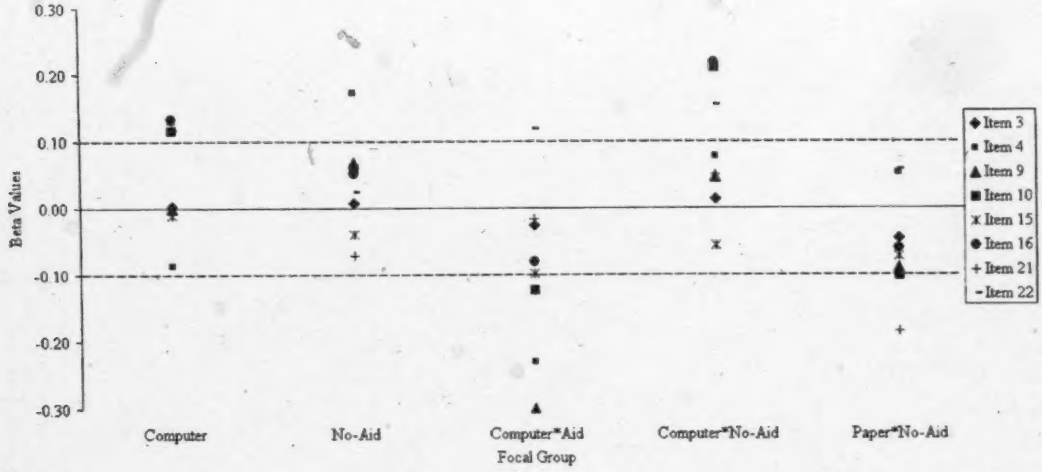


Figure 12: High Ability and No Valid Conclusion Syllogisms

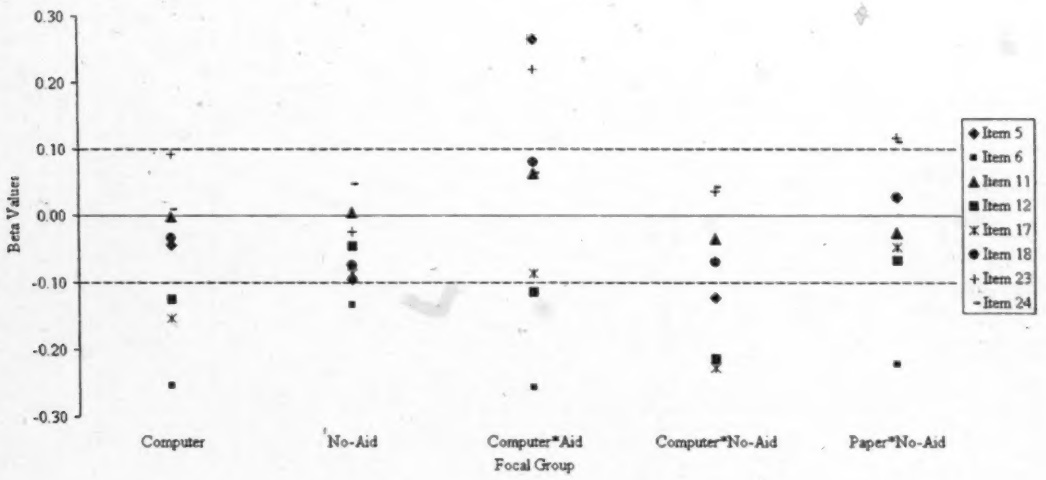


Figure 13: Memory Aid Use by Overall Ability and One-Model Syllogisms

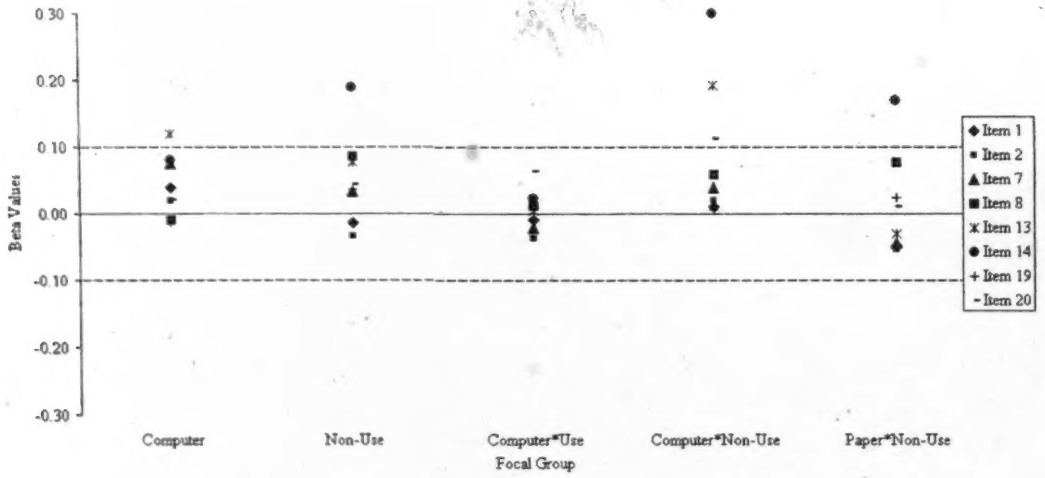


Figure 14: Memory Aid Use by Overall Ability and Multiple-Model Syllogisms

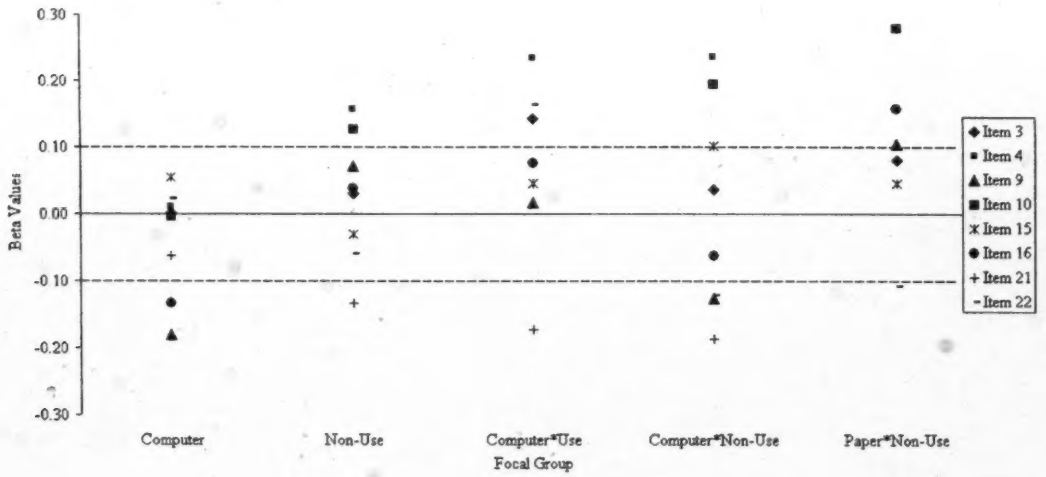


Figure 15: Memory Aid Use by Overall Ability and No Valid Conclusion Syllogisms

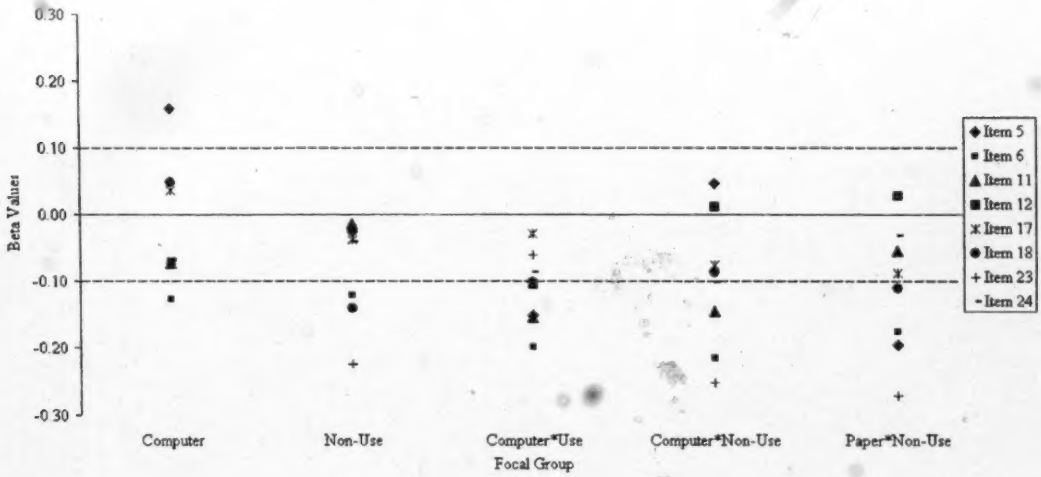


Figure 16: Memory Aid Use by Low Ability and One-Model Syllogisms

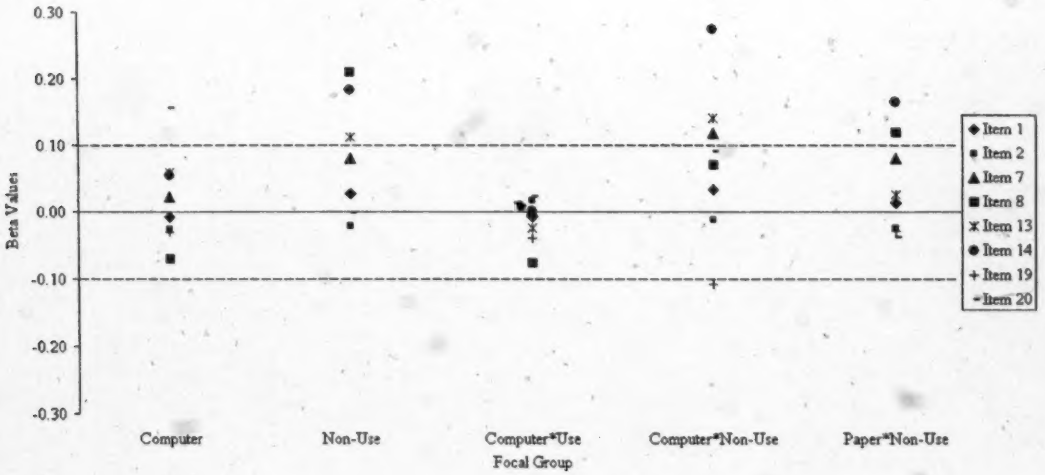


Figure 17: Memory Aid Use by Low Ability and Multiple-Model Syllogisms

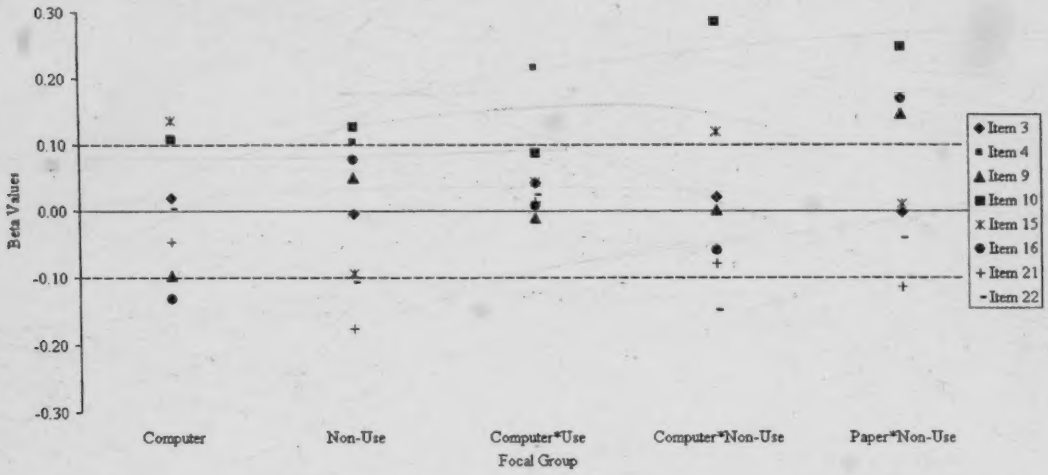


Figure 18: Memory Aid Use by Low Ability and No Valid Conclusion Syllogisms

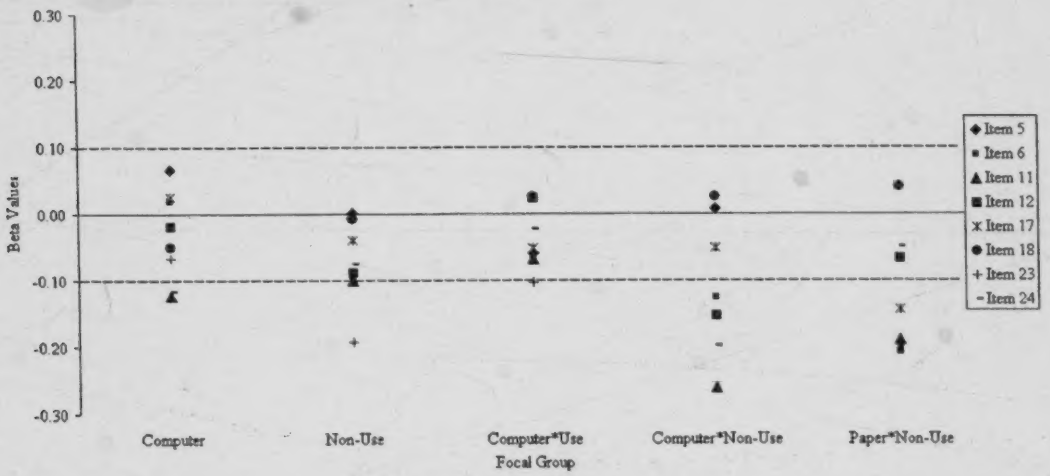


Figure 19: Memory Aid Use by High Ability and One-Model Syllogisms

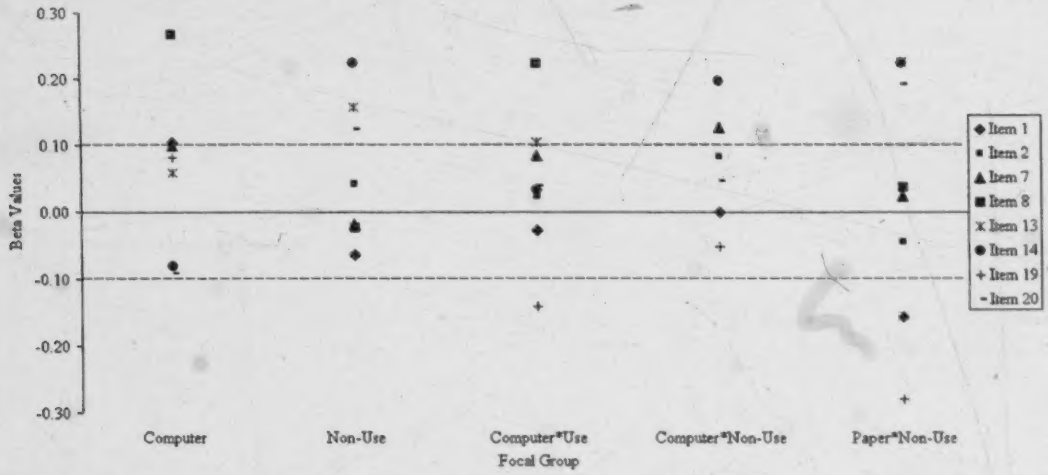


Figure 20: Memory Aid Use by High Ability and Multiple-Model Syllogisms

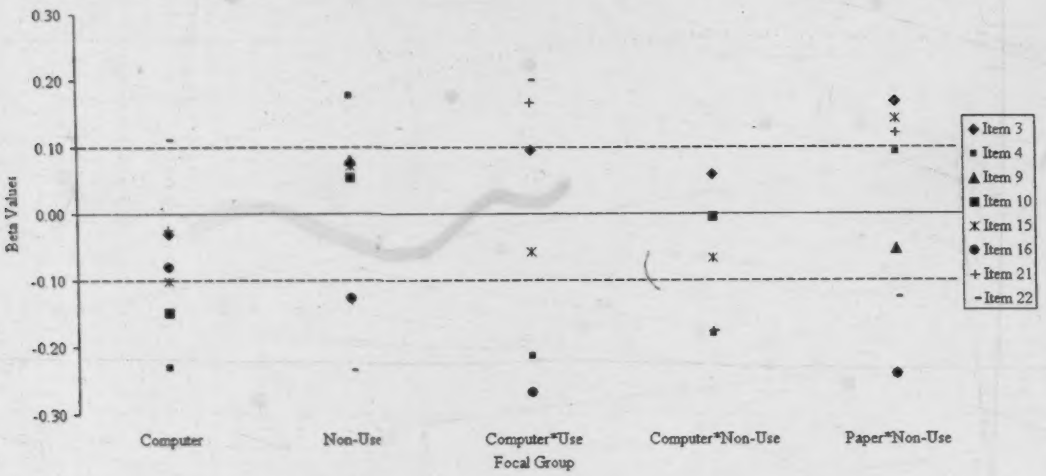
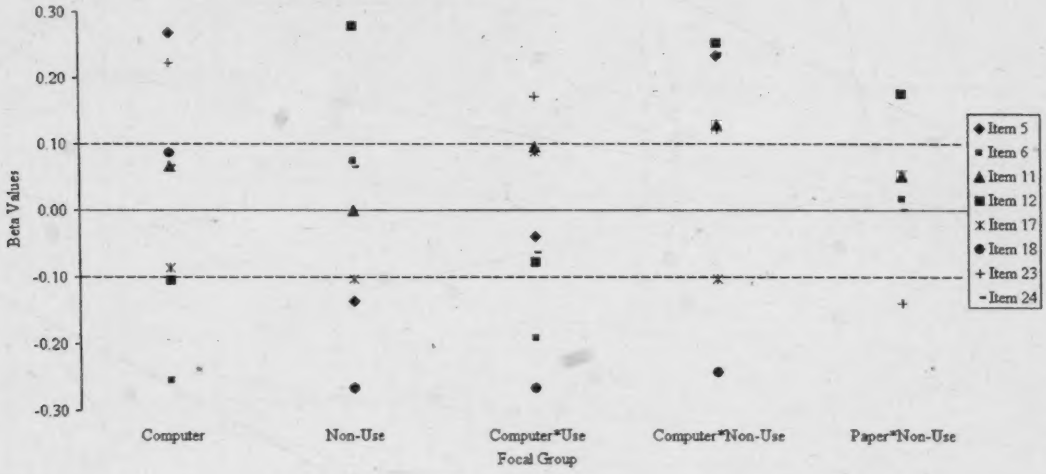


Figure 21: Memory Aid Use by High Ability and No Valid Conclusion Syllogisms



APPENDICES

Appendix A: Syllogisms

Appendix B: Paper Folding Test Example

Appendix C: Sample Instrument

Appendix D Item Characteristic Curves

Appendix A: Syllogisms

The following are the twenty-four syllogisms that appeared in the experimental instrument. Included with each are the response options made available to the examinees. The correct option is bolded for each item.

1) All A are B All B are C	**All A are C** Some A are C No A are C Some A are not C Nothing	7) All B are A All C are B	**All C are A** some C are A No C are A Some C are not A Nothing
2) Some A are B All B are C	All A are C **Some A are C** No A are C Some A are not C Nothing	8)No B are A All C are B	All C are A some C are A **No C are A** Some C are not A Nothing
3) No A are B Some B are C	All A are C Some A are C No A are C **Some A are not C** Nothing	9) No B are A Some C are B	All A are C some C are A No C are A **Some C are not A** Nothing
4) Some A are B No B are C	All A are C Some A are C No A are C **Some A are not C** Nothing	10) Some B are A No C are B	All A are C some A are C No A are C **Some A are not C** Nothing
5) Some A are B Some B are C	All A are C Some A are C No A are C Some A are not C **Nothing**	11) Some B are A Some C are B	All A are C some A are C No A are C Some A are not C **Nothing**
6) Some A are not B Some B are C	All A are C some A are C No A are C Some A are not C **Nothing**	12) Some B are not A Some C are B	All A are C some A are C No A are C Some A are not C **Nothing**

13) No A are B
All C are B

All A are C
some A are C
No A are C
Some A are not C
Nothing

14) All A are B
No C are B

All A are C
some A are C
No A are C
Some A are not C
Nothing

15) No A are B
Some C are B

All C are A
some C are A
No C are A
Some C are not A
Nothing

16) Some A are B
No C are B

All A are C
some A are C
No A are C
Some A are not C
Nothing

17) Some A are B
Some C are B

All A are C
some A are C
No A are C
Some A are not C
Nothing

18) No A are B
Some C are B

All A are C
some A are C
No A are C
Some A are not C
Nothing

19) Some B are A
All B are C

All A are C
Some A are C
No A are C
Some A are not C
Nothing

20) All B are A
Some B are C

All A are C
Some A are C
No A are C
Some A are not C
Nothing

21) No B are A
Some B are C

All C are A
some C are A
No C are A
Some C are not A
Nothing

22) Some B are A
No B are C

All A are C
some A are C
No A are C
Some A are not C
Nothing

23) Some B are A
Some B are C

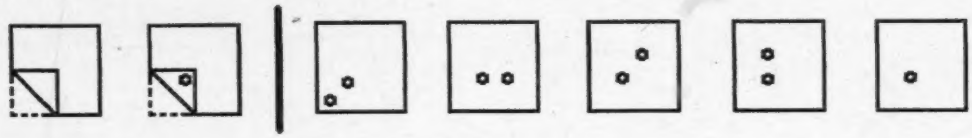
All A are C
some A are C
No A are C
Some A are not C
Nothing

24) Some B are not A
Some B are C

All A are C
some A are C
No A are C
Some A are not C
Nothing

Appendix B: Paper Folding Test Example

Figure B1: Paper Folding Test example



Appendix C: Sample Instrument

Figure C1: Sample instrument page one

Today you are going to take part in a study on people's problem solving ability. You will be asked to complete two sets of tasks as well as providing some demographic information. The first set consists of 24 problems. The second set consists of two pages of 20 problems each. You have a maximum of one hour to complete both sets. Please read all instructions before each set of tasks. If you have any questions at any time, do not hesitate to ask the researcher.

Before you begin the tasks, please fill out the information below:

What Faculty are you in?
a) Arts b) Sciences c) Other

What year of study are you in?
a) First b) Second c) Third d) Fourth e) Other

What is your age?
a) Under 18 b) 18-20 c) 21-24 d) 25-30 e) Over 30

What is your gender?
a) Female b) Male

Do you consider yourself a fluent English speaker?
a) Yes b) No

How comfortable do you feel working on computers?
a) Not at all b) Slightly c) Somewhat d) Comfortable e) Very

How would you assess your skill level in each of the following:	None	Poor	Fair	Good	Excellent
Word processing:	a	b	c	d	e
Email:	a	b	c	d	e
Games:	a	b	c	d	e
Searching the Web:	a	b	c	d	e

You may now begin the tasks by turning the page.

Figure C2: Sample instrument page two

In the following pages, you will see 24 pairs of statements about different people or groups of people, whom you should imagine as assembled in a room. After reading each pair of statements, please select the option that follows necessarily from these premises about the occupants of the room. If you consider that there is no conclusion that follows necessarily from the premises, please select "Nothing." Your conclusions should be based solely on the information presented in the statements, and NOT on plausible suppositions or general knowledge.

For example, please read the statements below:

1. All Artists are Beekeepers
   All Beekeepers are Chefs

a) All Artists are Chefs
b) Some Artists are Chefs
c) Some Artists are NOT Chefs
d) No Artists are Chefs
e) Nothing

From these statements, I can definitely conclude that All Artists are Chefs so I select the All Artists are Chefs option.

2. Some Acrobats are Bakers
   Some Bakers are Canoeists

a) All Acrobats are Canoeists
b) Some Acrobats are Canoeists
c) Some Acrobats are NOT Canoeists
d) No Acrobats are Canoeists
e) Nothing

From these statements, I can NOT conclude anything definite about acrobats and canoeists so I select the "nothing" option.

Please begin the task by turning the page. Do NOT skip any questions.

Figure C3: Sample instrument page three

Form A

1.	Some Managers are Singers Some Singers are Hairdressers	a) All Managers are Hairdressers b) Some Managers are Hairdressers c) Some Managers are not Hairdressers d) No Managers are Hairdressers e) Nothing
2.	Some Dancers are Doctors No Drivers are Dancers	a) All Doctors are Drivers b) Some Doctors are Drivers c) Some Doctors are not Drivers d) No Doctors are Drivers e) Nothing
3.	Some Gardeners are not Bookkeepers Some Bookkeepers are Musicians	a) All Gardeners are Musicians b) Some Gardeners are Musicians c) Some Gardeners are not Musicians d) No Gardeners are Musicians e) Nothing
4.	No Janitors are Swimmers All Plumbers are Janitors	a) All Plumbers are Swimmers b) Some Plumbers are Swimmers c) Some Plumbers are not Swimmers d) No Plumbers are Swimmers e) Nothing
5.	Some Inspectors are Authors Some Inspectors are Pipefitters	a) All Authors are Pipefitters b) Some Authors are Pipefitters c) Some Authors are not Pipefitters d) No Authors are Pipefitters e) Nothing
6.	Some Athletes are Nurses All Nurses are Secretaries	a) All Athletes are Secretaries b) Some Athletes are Secretaries c) Some Athletes are not Secretaries d) No Athletes are Secretaries e) Nothing
7.	All Golfers are Teachers All Carpenters are Golfers	a) All Carpenters are Teachers b) Some Carpenters are Teachers c) Some Carpenters are notTeachers d) No Carpenters are Teachers e) Nothing
8.	No Professors are Accountants Some Shoppers are Accountants	a) All Shoppers are Professors b) Some Shoppers are Professors c) Some Shoppers are not Professors d) No Shoppers are Professors e) Nothing
9.	No Engineers are Journalists Some Pharmacists are Journalists	a) All Engineers are Pharmacists b) Some Engineers are Pharmacists c) Some Engineers are not Pharmacists d) No Engineers are Pharmacists e) Nothing
10.	Some Psychologists are not Tailors Some Psychologists are Designers	a) All Tailors are Designers b) Some Tailors are Designers c) Some Tailors are not Designers d) No Tailors are Designers e) Nothing
11.	Some Farmers are Waiters Some Electricians are Farmers	a) All Electricians are Waiters b) Some Electricians are Waiters c) Some Electricians are not Waiters d) No Electricians are Waiters e) Nothing
12.	No Doctors are Dancers Some Dancers are Drivers	a) All Doctors are Drivers b) Some Doctors are Drivers c) Some Doctors are not Drivers d) No Doctors are Drivers e) Nothing

Figure C4: Sample instrument page four

Form A

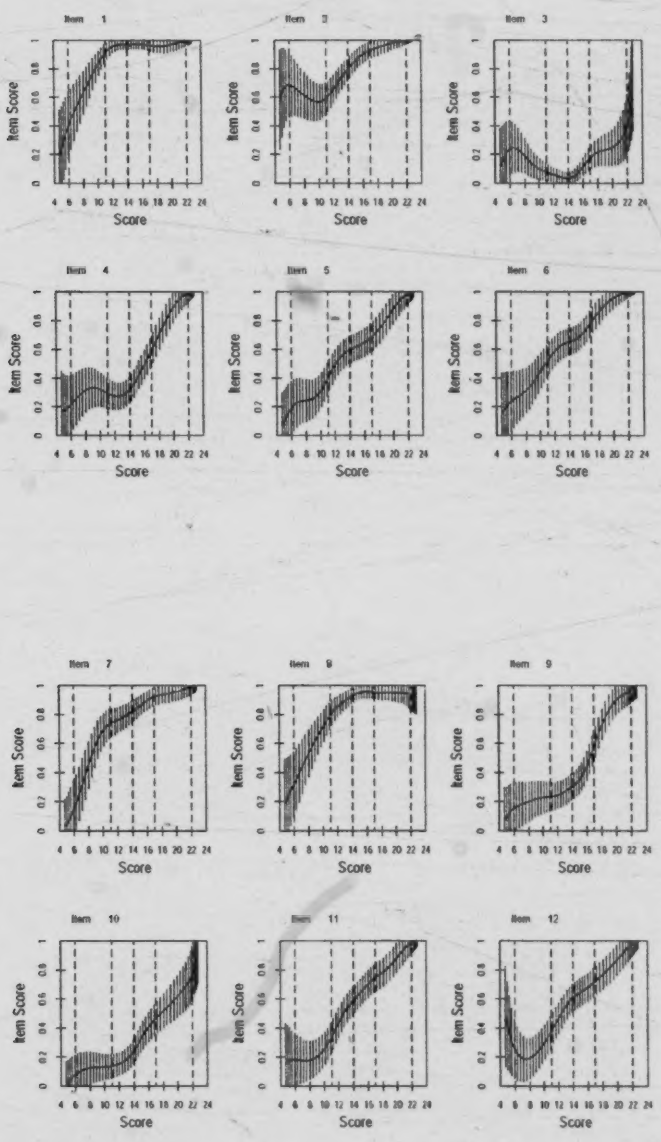
13. No Barbers are Biologists Some Barbers are Astronauts	a) All Astronauts are Biologists b) Some Astronauts are Biologists c) Some Astronauts are not Biologists d) No Astronauts are Biologists e) Nothing
14. Some Cabbies are Clerks No Cabbies are Couriers	a) All Clerks are Couriers b) Some Clerks are Couriers c) Some Clerks are not Couriers d) No Clerks are Couriers e) Nothing
15. Some Cooks are Bricklayers All Cooks are Firefighters	a) All Bricklayers are Firefighters b) Some Bricklayers are Firefighters c) Some Bricklayers are not Firefighters d) No Bricklayers are Firefighters e) Nothing
16. Some Architects are Chiropractors Some Dentists are Chiropractors	a) All Architects are Dentists b) Some Architects are Dentists c) Some Architects are not Dentists d) No Architects are Dentists e) Nothing
17. Some Surgeons are Researchers No Pilots are Researchers	a) All Surgeons are Pilots b) Some Surgeons are Pilots c) Some Surgeons are not Pilots d) No Surgeons are Pilots e) Nothing
18. No Welders are Entrepreneurs All Artists are Entrepreneurs	a) All Welders are Artists b) Some Welders are Artists c) Some Welders are not Artists d) No Welders are Artists e) Nothing
19. All Gardeners are Bookkeepers No Musicians are Bookkeepers	a) All Gardeners are Musicians b) Some Gardeners are Musicians c) Some Gardeners are not Musicians d) No Gardeners are Musicians e) Nothing
20. No Travellers are Librarians Some Counsellors are Travellers	a) All Counsellors are Librarians b) Some Counsellors are Librarians c) Some Counsellors are not Librarians d) No Counsellors are Librarians e) Nothing
21. Some Actors are not Photographers Some Inventors are Actors	a) All Inventors are Photographers b) Some Inventors are Photographers c) Some Inventors are not Photographers d) No Inventors are Photographers e) Nothing
22. Some Teachers are Golfers No Golfers are Carpenters	a) All Teachers are Carpenters b) Some Teachers are Carpenters c) Some Teachers are not Carpenters d) No Teachers are Carpenters e) Nothing
23. All Painters are Politicians All Politicians are Bankers	a) All Painters are Bankers b) Some Painters are Bankers c) Some Painters are not Bankers d) No Painters are Bankers e) Nothing
24. All Optometrists are Meteorologists Some Optometrists are Editors	a) All Meteorologists are Editors b) Some Meteorologists are Editors c) Some Meteorologists are not Editors d) No Meteorologists are Editors e) Nothing

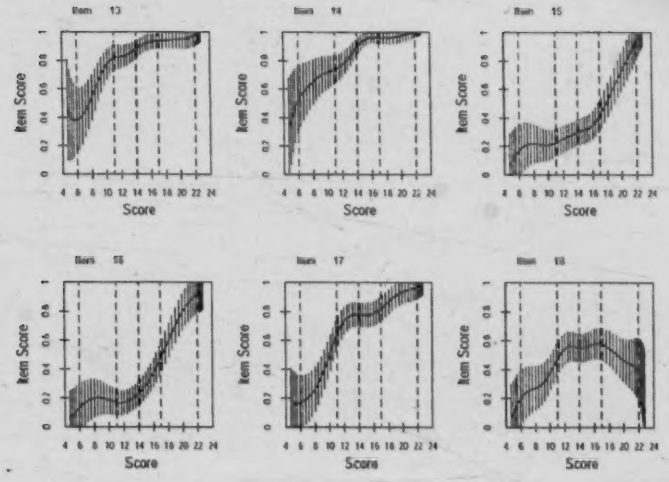
*Figure C5: S*ample instrument page eight

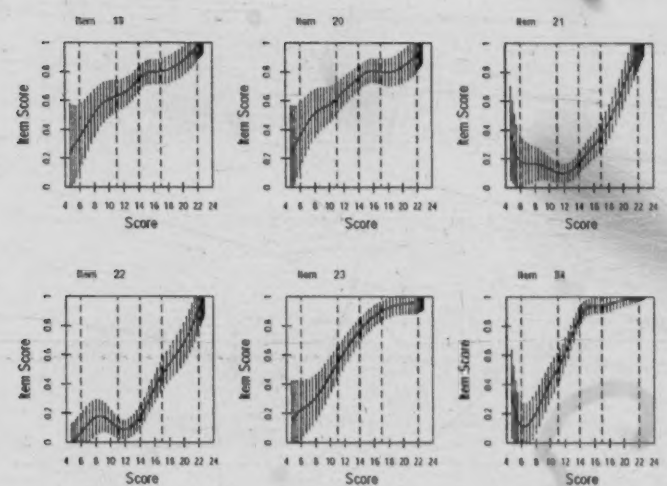
Thank you for your participation. You have now completed the SYL-124 research project. Please proceed to the researcher to receive your debriefing which contains information you will need for your final exam.

Appendix D: Item Characteristic Curves

Figure D1: Item Characteristic Curves (ICCs) for the twenty-four syllogisms




Item 13
Item 14
Item 15
Item 16
Item 17
Item 18
Item Score
Score


Item 19
Item 20
Item 21
Item 22
Item 23
Item 24
Item Score
Score